ADHO
Association for Digital Humanities Organization

DIGITAL HUMANITIES 2006

The First ADHO International Conference

Université Paris - Sorbonne

July 5th - July 9th
5 juillet - 9 juillet

Conference Abstracts
Résumé des communications

Centre de Recherche
Cultures Anglophones et
Technologies de l'Information

ISBN 2-9526916-0-6
Published by CATI, Université Paris-Sorbonne
Conference logo and cover design by Marie-Madeleine MARTINET.

Multimedia Information Retrieval

Tapio SEPPÄNEN
Department of Electrical and Information Engineering,University of Oulu, Finland

Massive amounts of digital data are being stored in networked databases for access via computers or mobile phone networks. Current ICT technology enables creation of digital data including such modalities as images, videos, multimedia, animations, graphics, speech, audio and text. A key problem will be how to organize and index this material to best support retrieval of relevant data. Not enough man-power is available to index or annotate all the terabytes of digital material by hand. Automatic and semi-automatic indexing mechanism should also be implemented. There are limits, however, to which extent it is sensible to index everything in the database. Content-based multimedia retrieval is a technology that applies digital signal processing and advanced pattern recognition techniques to analyse each potential object to determine its relevance. Applied techniques depend on the media type. For digital images, properties like texture, colour, and geometric structure primitives are analysed. For videos, movement of the image parts are tracked as well. For speech signals, automatic speech recognition, speaker recognition, gender recognition and language recognition can be used. Automatic detection of various audio events is also feasible today. A crucial problem is how to bridge the gap between the semantic concepts used by a user seeking information and the database descriptors automatically derived from the data. The difficulty lies in the design of reliably computable data features that describe the data in high-level terms such as semantic concepts. This can be viewed as a problem of mapping a query language to a data description language. In the presentation, examples of state-of-the-art data techniques are described that can be used for retrieving and analysing digital data.

Visualising 21st Century London

Robert TAVERNOR
Professor of Architecture and Urban Design, London School of Economics, UK

For centuries, artists have created memorable images of cities through drawings and paintings. These urban views - or vedute as Italians called them - are usually static, often idealised perspectival compositions. They depict key parts of a city - famous buildings and their settings - in ways that might be most favourably remembered by visitors: they were also used to impress rival cities. Then came photography, moving film and computerised animation. Yet, curiously, for all the advances in technology that were achieved during the late 20th century, static views - that collage together photographs and computer renderings of proposed buildings - are still being used when making major decisions about the visual image of 21st century London. Planning decisions about the positions, heights and clustering of large and tall buildings in London are being determined by experts assessing their likely visual impact on selected fixed viewing positions. Initially, ten principal "Strategic Views" were established across London in 1991, which were intended to safeguard the settings and silhouettes of St Paul's Cathedral and the Palace of Westminster. In 2004, The London Plan (2004), produced by the Mayor of the Greater London Assembly, refined and linked these as a View Protection Framework (2005) comprising "Panoramas, Townscape and Linear Views": this framework of views has been designed to protect a range of well-known images of the capital from surrounding hills and from the bridges crossing the River Thames. Now, whenever a major building is proposed in London, its visual impact on the existing cityscape is assessed through numerous photographs, taken from between 20 and 130 different viewing positions, into which "realistic" computer generated images of the proposed development are montaged. Buildings are consequently assessed - and composed - as sculptural objects in the urban landscape. This lecture will examine the visualizing methods being used and will assess the implications of composing a world capital through views.

References

Tavernor, R. (2001). 'A Computer Animation of Stanley Spencer's Church-House for the Stanley Spencer Exhibition, Tate Britain, London, April-June 2001.' In *Stanley Spencer.* Edited by Hyman, T.; Wright, P. Tate Publishing, 2001, pp.244-245.

Tavernor, R. (2002). Contemplating Perfection through Piero's Eyes. In *Body and Building. Essays on the Changing Relation of Body and Architecture,* (Edited Dodds, G and R. Tavernor), Cambridge (MA) and London: MIT Press, December 2001, HB. pp. 78-93.

Tavernor, R. (2004). 'From Townscape to Skyscape', *The Architectural Review,* March 2004, pp. 78-83.

Greater London Authority (2004), *The London Plan: download* at http://www.london.gov.uk/mayor/strategies/sds/index.jsp

Greater London Authority (2005), *London View Management Framework: download* at http://www.london.gov.uk/mayor/strategies/sds/spg-views.jsp

Weblink to the Cities Programme

http://www.lse.ac.uk/collections/cities/

Table of Contents

Plenary Sessions

Index of Abstracts

Single Sessions

Multiple Sessions

Posters

Single Sessions

The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic?

Dawn ARCHER
Department of Humanities, University of Central Lancashire

Andrea ERNST-GERLACH
Sebastian KEMPKEN
Thomas PILZ
Universität Duisburg-Essen

Paul RAYSON
Department of Computing, Lancaster University

1. Introduction

In this paper, we describe the approaches taken by two teams of researchers to the identification of spelling variants. Each team is working on a different language (English and German) but both are using historical texts from much the same time period (17th – 19th century). The approaches differ in a number of other respects, for example we can draw a distinction between two types of context rules: in the German system, context rules operate at the level of individual letters and represent constraints on candidate letter replacements or n-graphs; in the English system, contextual rules operate at the level of words and provide clues to detect real-word spelling variants i.e. 'then' used instead of 'than'. However, we noticed an overlap between the types of issues that we need to address for both English and German and also a similarity between the letter replacement patterns found in the two languages.

The aims of the research described in this paper are to compare manual and automatic techniques for the development of letter replacement heuristics in German and English, to determine the overlap between the heuristics and depending on the extent of the overlap, to assess whether it is possible to develop a generic spelling detection tool for Indo-European languages (of which English and German are examples).

As a starting point, we have manually-built letter replacement rules for English and German. We will compare these as a means of highlighting the similarity between them. We will describe machine learning approaches developed by the German team and apply them to manually-derived 'historical variant'-'modern equivalent' pairs (derived from existing corpora of English and German) to determine whether we can derive similar letter replacement heuristics. Using the manually-derived heuristics as a gold-standard we will evaluate the automatically derived rules.

Our prediction is that if the technique works in both languages, it would suggest that we are able to develop generic letter-replacement heuristics for the identification of historical variants for Indo-European languages.

2. German spelling variation

The interdisciplinary project "Rule based search in text databases with non-standard orthography" which is funded by the Deutsche Forschungsgemeinschaft [German Research Foundation] developed a rule-based fuzzy search-engine for historical texts (Pilz *et al*. 2005). Its aim of RSNSR is to provide means to perform reliable full text-search in documents written prior to the German unification of orthography in 1901.

On basis of around 4,000 manually collected one-to-one word mappings between non-standard and modern spellings, RSNSR follows three different paths to come up with an efficient rule set. Those are manual rule derivation, trained string edit distance and automatic rule learning. Additional mappings will be collected to further enhance the quality of those approaches. The manual derivation uses an alphabet of 62 different sequences, in parts historical *n*-graphs (e.g. <a>, <äu>, <eau>), built from combinations of the 30 standard graphemes of the German language. Being built manually, the alphabet considers linguistic restraints. Neither in context nor at the position of substitution non-lingual *n*-graphs (i.e. grapheme sequences that directly correspond to phonemes) are allowed. The context may also feature regular expressions using the *java.util.regex* formalism. The manually derived gold standard features the most elaborate rules. However the design of a rule set for the period from 1803 to 1806, based on only 338 *evidences*

took about three days to create. Furthermore, the manual derivation is prone to human-error. This is especially true as soon as the rule set exceeds certain limits where side effects become more and more likely.

The algorithm used to calculate the edit costs was proposed in 1975 by Bahl and Jelinek and taken up 1997 by Ristad and Yianilos who extended the approach by machine learning abilities. The authors applied the algorithm to the problem of learning the pronunciation of words in conversational speech (Ristad and Yianilos 1997). In a comparison between 13 different edit distances, Ristad and Yianilos' algorithm proofed to be the most efficient one. Its error rate on our list of *evidences* was 2.6 times lower than the standard Levenshtein distance measure and more than 6.7 times lower than Soundex (Kempken 2005).

The automatic generation of transformation rules uses triplets containing the contemporary words, their historic spelling variant and the collection frequency of the spelling variant. First, we compare the two words and determine so called 'rule cores'. We determine the necessary transformations for each training example and also identify the corresponding context. In a second step, we generate rule candidates that also consider the context information from the contemporary word. Finally, in the third step we select the useful rules by pruning the candidate set with a proprietary extension of the PRISM algorithm (Cendrowska 1987).

For this paper, we compared the German gold standard, mentioned above, with the two different machine learning algorithms. The string learning algorithm produces a fixed amount of single letter replacement probabilities. It is not yet possible to gather contextual information. Bi- or tri-graph operations are reflected by subsequent application of letter replacements. Therefore they do not map directly onto the manual rules. However, the four most frequent replacements, excluding identities, correspond to the four most frequently used rules. For the period from 1800 to 1806 these are *T→TH*, *Ä→AE*, *_→E* and *E→_*.

The manual and the automatic derived rules show obvious similarities, too. 12 of the 20 most frequently used rules from the automatic approach are also included in the manually built rules. For six other rules equivalent rules in the manual rule set exist. The rule *T→ET* from the automatic approach, for example, corresponds to the more generalised form *_→E* taken from the manual approach. And again do the first four rules match the four most frequent gold standard ones.

The automatic approaches, rule generation as well as edit distance, could be enhanced by a manual checking. Nevertheless, even a semi-automatic algorithm allows us to save time and resources. It is furthermore obvious, that the machine learning is already able to provide with a highly capable rule set for historical documents of German language.

3. English spelling variation

The existing English system called VARD (VARiant Detector) has three components. First, a list of 45,805 variant forms and their modern equivalents, built by hand. This provides a one-to-one mapping which VARD uses to insert a modern form alongside the historical variant which is preserved using an XML 'reg' tag. Secondly, a small set of contextual rules which take the form of templates of words and part-of-speech tags. The templates are applied to find real-word variants such as 'then' instead of 'than', 'doe' instead of 'do', 'bee' for 'be' and detection of the genitive when an apostrophe is missing. The third component consists of manually crafted letter replacement heuristics designed during the collection of the one-to-one mapping table and intended to reduce the manual overhead for detection of unseen variants in new corpora.

The rationale behind the VARD tool is to detect and normalise spelling variants to their modern equivalent in running text. This will enable techniques from corpus linguistics to be applied more accurately (Rayson et al., 2005). Techniques such as frequency profiling, concordancing, annotation and collocation extraction will not perform well with multiple variants of each word type in a corpus.

The English manual and automatically derived rules show a great deal of similarity. Nine of the twenty most frequent automatically derived rules are in the manual set. Eight other automatically derived rules have equivalents if we ignore context. Three automatically derived rules do not have a match in the manual version.

4. Conclusion

The motivation behind the two approaches of VARD and RSNSR differs. This reflects on the overall

structure of rules as well. While VARD is used to automatically normalise variants and thus takes more accurate aim to determine the correct modern equivalent, RSNSR focuses on finding and highlighting those historical spellings. Therefore its demands for precision are diminished while recall is the much more important factor. However, the approaches are highly capable of supporting each other and expanding their original field of application.

References

Cendrowska, J. (1987). PRISM: an algorithm for inducing modular rules. *Int. J Man-Machine Studies*, 27(4), pp.349-370.

Kempken, S. (2005). *Bewertung von historischen und regionalen Schreibvarianten mit Hilfe von Abstandsmaßen*. Diploma thesis. Universität Duisburg-Essen

Pilz, T., Luther, W., Ammon, U., Fuhr, N. (2005). Rule-based search in text databases with nonstandard orthography, *Proceedings ACH/ALLC 2005*, Victoria, 15 - 18 Jun 2005.

Rayson, P., Archer, D. and Smith, N. (2005) VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. In *proceedings of the Corpus Linguistics 2005 conference*, July 14-17, Birmingham, UK.

Ristad, E., Yianilos, P. (1997). Learning String Edit Distance, *IEEE Trans. PAMI*, 1997

Hessisches Staatsarchiv Darmstadt. http://www.stad.hessen.de/DigitalesArchiv/anfang.html (accessed 25 Nov. 2005)

Bibliotheca Augustana. FH Augsburg. http://www.fh-augsburg.de/~harsch/augustana.html (accessed 25 Nov. 2005)

documentArchiv.de http://www.documentarchiv.de (accessed 25 Nov. 2005)

A General Framework for Feature Identification

Neal AUDENAERT
Richard FURUTA
Eduardo URBINA

Center for the Study of Ditial Libraries, Texas A&M University

Large digital libraries typically contain collections of heterogeneous resources intended to be delivered to a variety of user communities. A key challenge for these libraries is providing tight integration between resources both within a single collection and across multiple collections. Traditionally, efforts at developing digital collections for the humanities have tended toward two extremes [5]. On one side are huge collections such as *the Making of America* [14, 15], *Project Gutenberg* [18], and *Christian Classics Ethereal Library* [13] that have minimal tagging, annotation or commentary. On the other side are smaller projects that closely resemble traditional approaches to editorial work in which editors carefully work with each page and line providing markup and metadata of extremely high quality and detail, mostly by hand. Projects at this end of the spectrum include *the William Blake Archive* [21], *the Canterbury Tales Project* [11], *the Rossetti Archive* [19], and *the Cervantes Project* [12]. These extremes force library designers to choose between large collections that provide an impoverished set of services to the collection's patrons on the one hand and relatively small, resource intensive projects on the other. Often, neither option is feasible.

An alternative approach to digital humanities projects recasts the role of the editor to focus on customizing and skillfully applying automated techniques, targeting limited resources for hand coding to those areas of the collection that merit special attention [6]. *The Perseus Project* [16] exemplifies this class of projects. Elucidating the internal structure of the digital resources by automatically identifying important features (e.g., names, places, dates, key phrases) is a key approach to aid in the development of these "middle ground" projects. Once identified, the internal structure can be

used to establish connections between the resources and to inform visualizations. This task is complicated by the heterogeneous nature of digital libraries and the diversity of user community needs.

To address this challenge we have developed a framework based approach to developing feature identification systems that allows decisions about details of document representation and features identification to be deferred to domain specific implementations of the framework. These deferred decisions include details of the semantics and syntax of markup, the types of metadata to be attached to documents, the types of features to be identified, the feature identification algorithms to be applied, and the determination of which features are to be indexed.

To achieve this generality, we represent a feature identification system as being composed of three layers, as diagramed in Figure 1. The core of the system is a "Feature Identification Framework" (FIF). This framework provides the major structural elements for working with documents, identifying features within documents, and building indices based on the identified features. Implementations customize components of the framework to interface with existing and new collections and to achieve domain specific functionality. Applications then use this framework, along with the appropriate set of customized modules, to implement visualizations, navigational linking strategies, and searching and filtering tools.

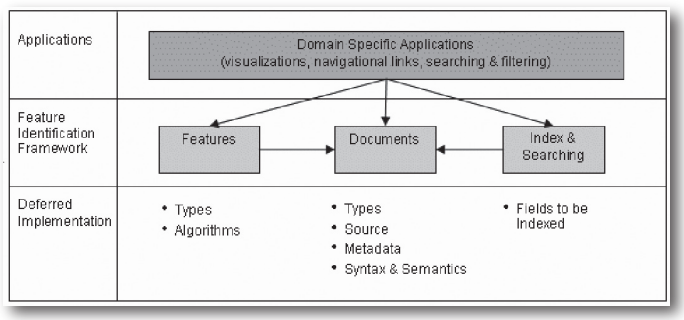


Figure 1: Three layered approach to designing a feature identification system

The document module implements the functionality needed to represent documents, manage storage and retrieval, provide an interface to searching mechanisms and facilitate automatic feature identification. It provides the following features:

1. Multiple types of documents (e.g., XML, PDF, RTF, HTML, etc) can be supported without modifying the APIs with which the rest of the system will interact.
2. Arbitrary syntactical constraints can be associated with a document and documents tested to ensure their validity. Notably, this helps to ensure that the markup of identified features does not violate syntactic or semantic constraints.
3. Metadata conforming to arbitrary metadata standards can be attached to documents.
4. Storage and retrieval mechanisms are provided that allow documents persistence to be managed either directly by the framework or by external systems.
5. Large documents can be broken into smaller "chunks" for both indexing and linking. Work in this area is ongoing.

The feature module builds on this base to provide the core toolset for identifying the internal structure of documents. Our design of this component reflects the highly contextualized nature of the feature identification task. The relevant features of a document can take many forms (e.g., a person or place, the greeting of a letter, a philosophical concept, or an argument against an idea) depending on both the type of document and the context in which that document is expected to be read. Equally contextualized are the algorithms used to identify features. Dictionary and statistically based methods are prevalent, though other techniques focusing on the semi-structured nature of specific documents have also yielded good results [3, 1, 4, 9, 2, 7]. Ultimately, which algorithm is selected will depend heavily on the choice of the corpus editor. Accordingly, our framework has been designed so that the only necessary property of a feature is that it can be identified within the text of a document and described within the structure provided by the document module.

For applications using the framework to effectively access and present the informational content, an indexing system is needed. Given the open ended nature of both document representation and the features to be identified, the indexing tools must inter-operate with the other customized components of the framework. We accomplish this, by utilizing adapters that are implemented while customizing the system. These adapters work with the

other customized components to specify the elements of each document to index.

To demonstrate and test this framework, we have implemented a prototype for a collection of official records pertaining Miguel de Cervantes Saavedra (1547-1616) originally assembled by Prof. Kris Sliwa [10]. This collection contains descriptions, summaries, and transcriptions in Spanish of nearly 1700 documents originally written from 1463 to 1681. These documents bear witness to the life of both Cervantes and his family and include inventory lists, birth and death certificates, and court testimonies.

Our application provides two primary points of access to the collection; a timeline navigator and a browsing interface. Following Crane, et al. [7], we have utilized proper names (people and places) and time as the two primary dimensions for structuring the documents in this collection. The timeline navigator, shown in Figure 2, displays a bar chart showing the distribution of the documents over time. Selecting a bar takes the reader to a more detailed view of the time period. Once the chart displays documents as single years, clicking on the bar for a single year brings up a display listing all documents from that year. The browsing interface, shown in Figure 3, allows readers to browse lists of both the people and the places identified within the collection. Upon selecting an item to view, a page presenting the resources available for that person or place is displayed. Currently, this includes a list of all documents in which the specified person has appeared and a bar graph of all documents in which that individual has been found as shown in Figure 4.

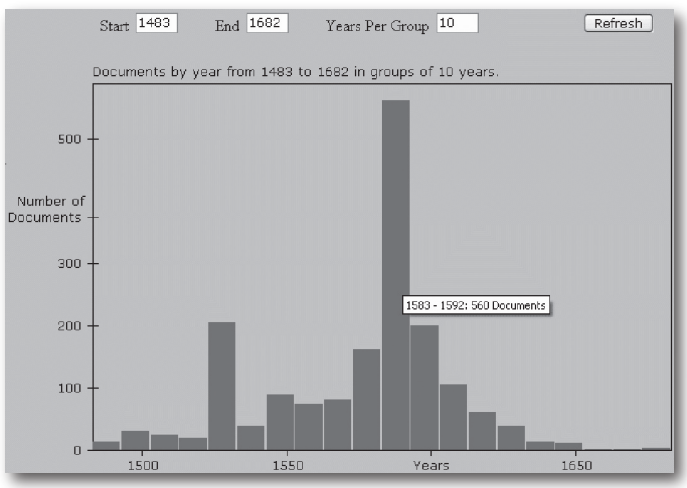


Figure 2: Timeline interface to the Sliwa collection

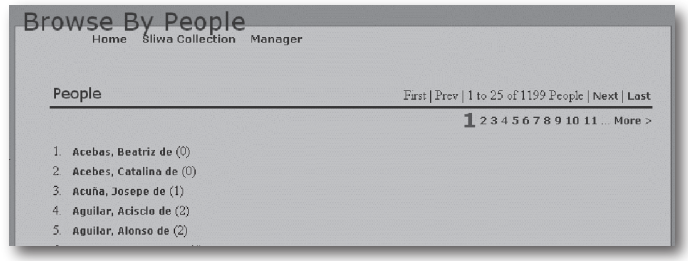


Figure 3: Browsing interface to the Sliwa Collection

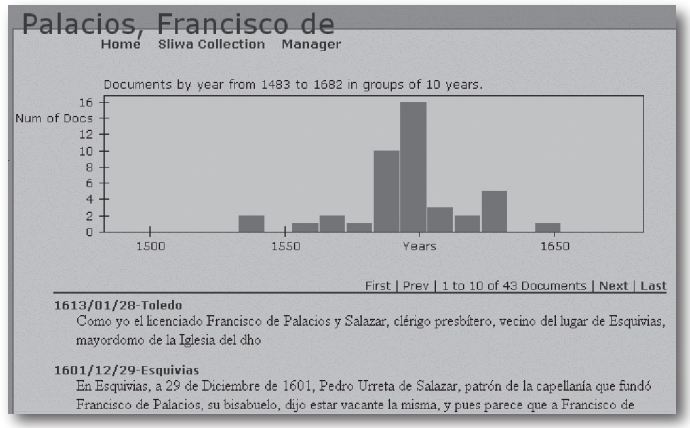


Figure 4: Browsing documents for Francisco de Palacios

Once the user has selected an individual document to view, through either the timeline or browsing interface, that document is presented with four types of features identified and highlighted. Identified people and places are used to automatically generate navigational links between documents and the pages presenting the resources for the people and places identified within a document. Dates and monetary units are identified and highlighted in the text.

One challenge with any framework based system is to ensure that the framework is not so general that customizing it requires more time and effort than writing an equivalent application from scratch. Our experience developing the Sliwa collection prototype suggests that our framework offers significant benefits. With the framework in place, we were able to develop and integrate new features in days; sometimes hours. Moreover, as sophisticated, general purpose features (e.g., pattern matching, grammatical parsers, georeferenced locations) are implemented, it becomes possible to customize and apply these features in new collections via a web-based interface with no additional coding involved. Custom document

formats are more complex to implement, but can serve in a wide variety of applications. The current implementation sufficient for most XML formats and work is underway to more fully support TEI encoded documents. Our approach provides strong support for the general components of a feature identification system thereby allowing individual projects to focus on details specific to the needs of particular collections and user communities.

We are currently working to apply this framework to a number of other projects, including diaries written during early Spanish expeditions into southern Texas [8], scholarly comments about the life and art of Picasso from *the Picasso Project* [17], and *the Stanford Encyclopedia of Philosophy* [20]. This will include further enhancements to the framework itself including support for feature identification that utilizes the structure of the document (including other identified features) in addition to the text and better support for accessing "chunks" within document in addition to the document as a whole. For the long term, we also plan to explore ways in which this framework can be used assist and shape editorial practices.

References

[1] **Bikel, D. M., R. Schwartz, and R. M. Weischedel,** 1999. *An Algorithm that Learns What's in a Name*. *Machine Learning,* 34(1-3): p.211-231.

[2] **Callan J., and T. Mitamura.** 2002. Knowledge-based extraction of named entities. In *Proceedings of the eleventh international conference on Information and knowledge management*. McLean, Virginia, USA: ACM Press

[3] **Chinchor, N. A**. 1998. Overview of MUC-7/MET-2. *In Proceedings of the Seventh Message Understanding Conference* (MUC-7). Fairfax, Virginia USA. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

[4] **Cohen, W. W., and S. Sarawagi**. 2004. Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods, In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*. Seattle, WA, USA: ACM Press

[5] **Crane, G.** 2000. Designing documents to enhance the performance of digital libraries: Time, space, people and a digital library of London. *D-Lib Magazine,* 6(7/8).

[6] **Crane, G., and J. A.** 2000. Rydberg-Cox. New technology and new roles: the need for "corpus editors." *In Proceedings of the fifth ACM conference on Digital libraries*. San Antonio, TX USA: ACM Press.

[7] **Crane, G., D. A. Smith and Wulfman, C. E.** 2001. Building a hypertextual digital library in the humanities: a case study on London. In *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries*. Roanoke, VA USA: ACM Press.

[8] **Imhoff, B.**, ed. 2002. *The diary of Juan Dominguez de Mendoza's expedition into Texas (1683-1684): A critical edition of the Spanish text with facsimile reproductions*. Dallas, TX: William P. Clements Center for Southwest Studies, Southern Methodist University.

[9] **Mikheev, A, M. Moens and C Grover**. 1999. Named Entity recognition without gazetteers, In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Bergen, Norway: Association for Computational Linguistics.

[10] **Sliwa, K.**. 2000 *Documentos Cervantinos: Nueva recopilación; lista e índices*. New York: Peter Lang.

[11] "The Canterbury Tales Project," De Montfort University, Leicester, England. http://www.cta.dmu.ac.uk/projects/ctp/index.html. Accessed on May 25, 2002.

[12] "The Cervantes Project." E. Urbina, ed. Center for the Study of Digital Libraries, Texas A&M University. http://csdl.tamu.edu/cervantes. Accessed on Feb 7, 2005.

[13] "Christian Classics Ethereal Library", Calvin College, Grand Rapids, MI. httphttp://www.ccel.org/. Accessed on Sept 8, 2005.

[14] "Making of America." University of Michigan

http://www.hti.umich.edu/m/moagrp/. Accessed on Sept 8, 2005.

[15] "Making of America." Cornell University. http://moa.cit.cornell.edu/moa/. Accessed on Sept 8, 2005.

[16] "Perseus Project" G. Crane, ed. Tufts University. http://www.perseus.tufts.edu/. Accessed on Sept 9, 2005.

[17] "The Picasso Project", E. Mallen, ed. Hispanic Studies Department, Texas A&M University. http://www.tamu.edu/mocl/picasso/. Accessed on Feb 7, 2005.

[18] "Project Gutenberg." Project Gutenberg Literary Archive Foundation. http://www.gutenberg.org/. Accessed on Sept 9, 2005.

[19] "The Rossetti Archive." J. McGann, ed. The Institute for Advanced Technologies in the Humanities, University of Virginia. http://www.rossettiarchive.org/. Accessed on Feb 7, 2005.

[20] "Stanford Encyclopedia of Philosophy." Stanford University. http://plato.stanford.edu/. Accessed on Nov 14, 2005.

[21] "The William Blake Archive" M. Eaves, R. Essick, and J. Viscomi, eds. The Institute for Advanced Technology in the Humanities. http://www.blakearchive.org/. Accessed on Sept 9, 2005.

Demand Balancing in Fee-free Resource Sharing Community Networks with Unequal Resource Distribution

Edmund BALNAVES
University of Sydney
Prosentient Systems Pty Ltd
ejb@it.usyd.edu.au

1. Introduction

Even the largest libraries struggle to maintain a comprehensive journal collection. In 2003 Australian universities subscribed to over 1,300,000 journals of which of which 974,000 were in aggregate digital collections. This represented over 273,000 new serial titles and over 150,000 cancellations (Council of Australian University Libraries, 2005). One emerging approach is for libraries to form consortia that take a joint subscription to digital resource collections. The consolidation of substantial collections with direct delivery has seen the gradual attrition of subscriptions to the traditional print format (Fox & Marchionini, 1998; Weiderhold, 1995), but this cost saving is offset by the substantial increase in digital resources. The wealth of international research resources presents an even greater dilemma for small research institutions: how to effectively and economically access such a wide base of information resources within sometimes highly constrained budgets. The cost reductions obtained through aggregate subscriptions and consortia do not necessarily offset the net growth of fee-for-use published resources, and may have the consequence of centralizing subscriptions through a few large distributors – with the long-term collection risk that this centralization presents.

Small research libraries that cannot afford participation in national inter-library loan networks have formed fee-free networks of collaborating libraries that share their journal resources. While a fee-free Inter-Library Loan (ILL) service offers obvious attractions to smaller

participating libraries, alternative economic approaches are needed to avoid excessive demand on resource-rich members, and to avoid the phenomenon of "free-riders". This paper presents the resource distribution approaches that have been used to balance resource demand in GratisNet, an Australian network of 250+ health research libraries, where collaboration is fee-free but resource holdings among member are unequal. Dynamic ranking resource-based approaches are used to encourage the equitable distribution of resource load.

2. Economics of demand balancing in a fee-free network

Hooke (1999) highlighted the need for evidence -based approaches in the management of information services. In a fee-based environment, the metrics for efficiency may centre of cost versus speed of supply. Fee-free collaboration does not have the same economic driver for equilibrium between demand and resource supply that emerges in the long term in a fee-based service. Furthermore, resource sharing networks operating in a fee-free environment face several risks that are common to voluntary online communities. In Gaming Theory "outcomes" and "payoffs" are differentiated (Shubik, 1975). In the case of ILL collaboration, the payoffs are the supply of particular ILL requests in exchange for the provision of requests raised by other libraries at the risk of absorbing the costs of supplying requests raised by other libraries. The outcomes include access to a wider base of research resources than would otherwise be available to the library, and the potential for requests to exceed loans and constraints on the limit of demands based on membership of a closed community. One of the risks is the "free-rider" phenomenon, or those who take the benefit of membership of a collaborating community but provide no net contribution of resources. "Free-riders" can be managed in a number of ways: through "closed shops" (the example of unions that limit benefits to those who are members only), or through adjustment of the payoffs(Hamburger, 1979).

3. Demand balancing

Imbalanced distribution of workload in a fee-free environment, if unmanaged, can create imperfect resource management through inequitable distribution of demand over time. These imperfections may be expressed in terms of a reluctance to declare resources or supply requests (a form of compliance failure), or through inequitable distribution of demand resulting in a delay in supply (a queuing problem). This reduces the payoff potential for the larger members of the network. While there is a risk that libraries may reduce their own collections through reliance on wider networks, the trade-off is the delay in fulfilling requests when they are completed through an ILL rather than directly out of their own collection.

The rational choice in fulfilment of an individual ILL request in a fee-free environment is the selection of the nearest library that has the highest probability of fulfilling the request. This provides the best payoff in probability of fulfilment and timeliness of supply. However, aggregated over time this choice is likely to place a larger burden on those participating libraries with the largest collection of resources in a given region. Where staff represents as much as 80% of document supply cost (Morris, 2004), this can be a considerable burden on larger participating libraries. In the GratisNet network, search results for resources held by members of the GratisNet network are inversely ranked based on historical workload contribution. Participating libraries are requested to select from resources in the top-ranked selections presented, but compliance is voluntary. Participating libraries supply ILL requests at no charge to members and with no specific reciprocity. The objective of the ranking process is to adjust the payoff implied by a ration selection of the largest, nearest library by tempering this choice through ranking of search results based on previous workload of participating libraries. Libraries with a higher historical workload are ranked lower in search results. Libraries are encouraged to select from one of the first three listed libraries that have holdings in the journal they are requesting.

Table 1 shows the percentage of libraries that by-passed the computer-recommended ranking when raising ILL requests for the years 2002, 2003 and 2004. Since compliance is voluntary, this change demonstrates increasing trust in the workload distribution mechanisms. While participating libraries do exercise a measure of discretion in selecting outside the recommended rankings, voluntary compliance to the ranking recommendations is generally good and has improved over time.

	Bypass %
2002	26,07
2003	16,40
2004	10,66

Table 1 Compliance in ranking selection

Game-theoretic formulations can provide a useful approach to the design of co-operative IT systems(Mahajan, Rodrig, Wetherall, & Zahorjan, 2004). To illustrate the contrast between a time-efficient system for ILL delivery and one which distributes workload across the network, these same transactions were reprocessed under to a game scenario which simulated a rational select on the basis of proximity and breadth of holdings matching the request for the most recent two years. The objective of this scenario was to contrast the aggregate effect of load-based ranking with a utility-based approach to request ulfilment (see Table 2 below). In a time-efficient approach, larger libraries are consistently net providers, reducing their aggregate payoff from participation. Pure Egalitarianism takes the approach over time that yields the highest combined utility to participating libraries. The voluntary element of the ranking yields a "relative egalitarianism" which balances the result that yields utility achieved overall with the lowest level of frustration.(Moulin, 1988).

	2003	2004
Transactions	150155	128365
Distribution of transactions to the top 20 largest libraries (demand-balanced)	39003	33212
Distribution of transactions to the top 20 largest libraries (utility-based)	56815	46760
Distribution of transactions to the 20 smallest libraries (demand-balanced)	1454	1154
Distribution of transactions to the 20 smallest libraries (utility-based)	779	754

Table 2 Contrasting load-based ranking to utility-based ranking

The risk facing groups collaborating on a fee-free basis is that inequity of resource distribution could result in the resignation of members where their level of "frustration" exceeds the benefit they gain from participation.

4. Conclusion

Participating libraries in the GratisNet network commit to supplying ILL requests at no charge and with no specific reciprocity, on the basis that they can be confident that an increase in demand on their library will be balanced progressively with a lower ranking in search results. Transactions for the period 2003 to 2005 are analysed to illustrate the ways in which a ranking-based approach to resource discovery improves workload distribution for participating members overall. Results from the GratisNet network illustrate the effectiveness of formal approaches to resource distribution in fee-free collaborative networks. This analysis also gives an insight into the ways in which service metrics can help in the management of workload in a fee-free environment.

References

Council of Australian University Libraries. (2005). *Caul statistics 2003. Accessed 29-nov-2005 from http://www.Caul.Edu.Au/stats/caul2003-pub.Xls.*

Fox, E. A., & Marchionini, G. (1998). Toward a worldwide digital library. *Communications of the ACM,* 41(4), 29-32.

Hamburger, H. (1979). *Games as models of social phenomena*. San Francisco: Freeman.

Hooke, J. (1999). Evidence-based pracitice and its relevance to library and information services. *LASIE*(Sept 1999), 23-34.

Mahajan, R., Rodrig, M., Wetherall, D., & Zahorjan, J. (2004). *Experiences applying game theory to system design*. Paper presented at the ACM SIGCOMM '04 Workshop, Aug 30-Sep 3, 2004, Portland, Oregan, USA.

Morris, L. R. (2004). How to lower your interlibrary loan and document delivery costs: An editorial.

Journal of Interlibrary Loan, Document Delivery & Information Supply, 14(4), 1-3.

Moulin, H. (1988). *Axioms of cooperative decision making*. Cambridge: Cambridge University Press.

Shubik, M. (1975). *The uses and methods of gaming*. NY: Elsevier.

Weiderhold, G. (1995). Digital libraries, value and productivity. *Communications of the ACM*, 38(4), 85-97.

TEI P5: What's in It for Me?

Syd BAUMAN
Women Writers Project, Brown University

Lou BURNARD
Oxford University Computing Services, Oxford University

The Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange are a community based standard for text encoding that aim to "apply to texts in any natural language, of any date, in any literary genre or text type, without restriction on form or content." The basic idea is not to tell scholars, librarians, and other encoders *what* to encode, but rather *how* to encode that which they choose to record. Of course the Guidelines cannot possibly anticipate every feature that users may wish to encode, and therefore the capability to extend the encoding scheme described by the Guidelines in a consistent, easily understood, and interchangeable manner is paramount.

Over the past few years the Text Encoding Initiative Consortium (the organization charged with maintaining, developing, promulgating, and promoting the Guidelines) has been working steadily toward a new release of the Guidelines. This much anticipated version, referred to as "P5", is significantly different from the current Guidelines ("P4"), and yet performs the same basic function of providing a community based standard for encoding literary and linguistic texts.

In this presentation, the TEI editors will present P5 as of the latest release, with an emphasis on the ease with which P5 can be customized to particular uses. The talk will start with an overview of what P5 is, what it is good for, and why one would want to use it, and then progress to some of the detailed differences between P4 and P5.

Topics addressed will include:

* The general goal of the TEI Guidelines
  - TEI is a community initiative, driven by the needs of its members and users

* How work gets done in the TEI - technical council and work groups
  - open source using Sourceforge
  - special interest groups
* Why do this -- isn't P4 good enough?
  - P4 is just P3 in and using XML
  - a lot has happened since P3 was released, including the creation of the W3C and the acceptance of Unicode
  - there are arenas P4 does not cover
  - lots of improvements, repairs, etc. are in order
* What's new and different
  - infrastructural
  + schemata
  + datatypes
  + classes
  + customization
  - attributes used wherever textual content allowed
  - major updates
  + manuscript description
  + character sets & language identification
  + feature structures (now an ISO standard)
  + pointing mechanism
  - less major updates
  + dictionaries
  + multiple hierarchies
  + support for graphics and multimedia
  + support for stand-off markup
  - new updates
  + personography
  + terminological databases
  + collation sequences
* customization
  - customizations permit a project to select which parts of the TEI scheme they will use, and to add new bits if needed
  - in P5, all uses of TEI are customizations of one sort or another
  - customizations, and the user's documentation for them, are written in a TEI file
  - thus customizations themselves can be interchanged or even shared
  - in theory, a customization can use another customization as its starting point
  - thus permitting customizations of customizations
* The TEI universe
  - the TEI universe is one where all projects share a common base, but many use additional, local markup constructs
  - clusters of similar projects can share common subsets of additional markup constructs

Méthodes automatiques de datation d'un texte

Michel BERNARD

Centre de recherche Hubert de Phalèse. Université Sorbonne Nouvelle

Il est classique, en philologie, de dater un texte en utilisant comme « terminus post quem » la date d'attestation de certaines de ses formes. L'utilisation de grandes bases de données comme Frantext permet d'envisager l'automatisation de cette opération, et par conséquent son application à toutes les formes d'un texte, ce qui augmente la précision de la datation. Un certain nombre de précautions (choix du corpus de référence, prise en compte des variations graphiques,...) doivent être prises pour mener à bien cette opération, qui peut être améliorée par la prise en compte, au-delà de la seule date de première attestation, de la fréquence d'utilisation, tout au long de l'histoire de la langue, des formes du texte à dater. _La datation du vocabulaire d'un texte permet également d'apprécier le degré d'archaïsme et de néologie mis en oeuvre par l'auteur, ce qui a des applications en stylistique et en histoire des genres. _On envisagera aussi l'utilisation de dictionnaires d'attestations pour effectuer ces opérations de datation, dictionnaires existants ou à constituer dans ce but. _Les exemples porteront aussi bien, pour validation, sur des textes dont la date est connue de l'histoire littéraire que de textes dont la datation fait aujourd'hui l'objet de controverses.

It is traditional, in philology, to date a text by using as a "terminus post quem" the date of first attestation of certain words. The use of a great data base as Frantext makes it possible to envisage the automation of this operation, and consequently its application to all the words of a text, which would increase the precision of the dating. A certain number of precautions (choice of the reference corpus, recognition of the graphic variations...) must be taken to perform this operation, which can be improved by the taking into account, beyond the only date of first attestation, of the frequency of use, throughout the history of the language, of the words of the text to be dated. The dating of the vocabulary of a text also makes it possible to appreciate the degree of archaism and neology implemented by the author, which has applications in stylistics and history of the genres. One will consider also the use of dictionaries of attestations to carry out these operations of dating, dictionaries existing or to be constituted to this end. The examples will relate as well, for validation, to texts whose date is known in literary history and to texts whose dating is the subject of controversies.

Critique de la bibliométrie comme outil d'évaluation; vers une approche qualitative.

Marc BERTIN
Jean-Pierre DESCLÉS
Yordan KRUSHKOV

Laboratoire LaLICC (Langage, Logique, Informatique, Cognition et Communication), Université Paris-Sorbonne/CNRS UMR8139

The evaluation of the researchers is a problem we are facing at present times. Bibliometric methods use mainly statistical tools, consequently, this approach does not provide any tools of qualitative evaluation. Bibliometry is a quantitative evaluation of literature. Numerous studies contributed to the advance of this science and to the discovery of indicators allowing to estimate the productivity of a researcher, a country or an institution. So, we suggest to examine the specific details of the bibliographic references in the texts. Our hypothesis is that by locating the indicators in corpus we will provide sentences localization with linguistics clues. These specific linguistic units, using the method of contextual exploration, give us opportunities to annotate articles with information about citation. We uses the contextual exploration method, a linguistic approach, which allows us to annotate automatically the text. Contextual Exploration, proposed and developed by Jean-Pierre Desclés and LaLICC group, is based upon the observation that it is possible to identify specific semantic information contained in certain parts of text. This approach does not depend on the specific domain of articles. Furthermore, we will resolve the limitation of statistical method, related to a possible distortion resulting from the negative citation of authors.The informatic application of this study will be integrated into the platform EXCOM (EXploration COntexuel Multilingue). We demonstrated that it was possible to identify and to annotate the textual segments from bibliography. Furthermore, through this linguistic study of the textual segments, we identified and categorized linguistics clues. The annotation of these linguistics clues facilitates automated processing within the frame of the contextual exploration method implemented in the platform EXCOM. This phase allows us to qualify the relations between the author, the coauthors and also the bibliography. It is possible to classify citation according to a qualitative approach and it is offering a better use of the bibliography.

1. Les outils bibliométriques

Si l'évaluation de la science et de la production scientifique des chercheurs est un débat récurrent, il est de plus en plus présent ces dernières années. L'approche bibliométrique est la plus développée pour ne pas dire la seule. On pourra citer les écrits de Bradford, Lotka, Zipf et des travaux portant sur l'unification de ces lois. Nous pouvons également présenter les calculs des distributions à travers les mesures de concentration ou l'entropie de Shannon. Ces méthodes sont principalement issues de l'univers des statistiques et des grands nombres. Au-delà de l'aspect théorique des distributions bibliométriques, nous pourrions citer un homme, Eugene Garfield, qui à travers son article « Citation indexes of science : a new dimension in documentation throught association of ideas » a proposer un outil dévaluation de la science connu sous le nom de Facteur d'Impact. Afin de montrer l'importance de disposer d'un tel outil, nous présenterons deux catégories d'indicateur et leur utilité: les indicateurs univariés et les indicateurs relationnels.

Les indicateurs univariés permettent avant tout d'évaluer la productivité d'un chercheur, d'un laboratoire, d'un domaine ou bien d'un pays. Cependant, cet indicateur reste déconseillé au niveau de l'individu. Les quantités mesurables peuvent être le nombre de publications, le nombre de co-signatures et co-publications, le nombre de citations qui montrent l'impact des articles cités. Les liens scientifiques des citations montrent le rapport d'influence entre communautés scientifiques. Nous retrouvons ici le fameux facteur d'impact proposé par M. Garfield et qui est utilisé par l'Information Science Institute. Enfin, le nombre de brevets ainsi que les citations de brevets sont également des indicateurs d'inventivité, d'innovation et de capacité technologique montrant le résultat des ressources investies dans les activités de recherche et développement.

Les indicateurs relationnels sont principalement les co-citations et les co-occurrences de mots. Les co-citations

présentent les réseaux thématiques et l'influence des auteurs. L'indice d'affinité mesure les liens entre les pays en calculant le taux relatif des échanges scientifiques entre deux pays de masse scientifique comparable, pendant une période donnée par rapport à l'ensemble de la coopération nationale de ces deux pays. L'utilisation des citations ou des co-auteurs permet de proposer des relations entre auteurs sous forme matricielle afin d'obtenir des réseaux.

2. Les limitations de l'outil

La bibliométrie apporte donc une mesure des activités de recherche, mais un ensemble de limites d'ordre technique et conceptuel ne permet pas l'utilisation à l'unanimité des indicateurs correspondants. Nous relèverons les biais suivants: actuellement, seul le premier auteur est pris en compte. Il faut également considérer les fautes de frappe et l'homonymie. Les domaines sont inégalement représentés et les indicateurs s'appliquent très difficilement pour les sciences humaines et sociales. Toutes les revues ne sont pas recensées et pour celles qui le sont, il peut y avoir sur ou sous-estimation de la revue et donc des travaux et des équipes. On notera que l'autocitation ou la citation d'un article controversé n'est pas abordé par l'approche statistique. De plus, les ouvrages ne sont pas pris en compte. Nous pouvons aussi constater que deux ans ne suffisent pas pour qu'un article se révèle or il s'agit de la durée retenue pour le calcul du facteur d'impact.

Si les journaux en science de l'information s'intéressent naturellement à cette problématique, nous constaterons que cette question touche des domaines qui dépassent ce cadre. Garfield mentionnera par la suite : « I first mentioned the idea of an impact factor in 1955. At that time it did not occur to me that it would one day become the subject of widespread controversy. [...] I expected that it would be used constructively while recognizing that in the wrong hands it might be abused ». On peut non seulement affirmer que les moyens actuels ne permettent pas d'identifier la valeur d'un papier. Mais qu'elle conduit à des pratiques qui peuvent mettre en péril la qualité des articles. Les conséquences ne sont pas sans importance. Cela peut provoquer des comportements antiscientifiques comme le plagiat, la publication dans une revue où le FI est élevé plutôt que dans une revue adéquate ou bien encore de diviser les données en partie ridiculement petites. Nous sommes dans l'ère du « publier ou mourir ». De ce déclin de la diversité, nous risquons d'avoir à moyen terme une recherche homogène. Si les articles pointant du doigt les biais introduits par cette méthode d'évaluation sont de plus en plus nombreux, ils ne proposent cependant guère de solutions innovantes, seulement de nouvelles approches statistiques permettant de minimiser les biais introduits.

Il est difficile de mesurer la qualité d'une production scientifique car la bibliométrie, et plus spécifiquement les indicateurs, caractérisent le contenant et non le contenu. Ils sont des mesures et non des signes précieux de la qualité de la recherche. Si la bibliométrie rajoute de la valeur à la vue des pairs, elle ne peut que difficilement les remplacer. Ce débordement, en dehors des canaux de communication classique n'est-il pas le signe précurseur que l'hégémonie du FI de ces cinquante dernières années a vécu et qu'il est désormais nécessaire de passer d'une évaluation quantitative à une évaluation qualitative de la publication scientifique.

3. Une approche qualitative

Une nouvelle approche de cette problématique doit être envisagée. Nous devons désormais nous intéresser à l'auteur, à ces co-auteurs et également au contenu d'un article selon une approche qualitative. Pour cela, une réflexion sur l'étude des publications doit être entreprise. Sans prétendre fournir un traitement sémantique complet d'un article scientifique, nous pourrons dans un premier temps considérer les relations sémantiques entre l'auteur, les co-auteur et les références bibliographiques. Il serait tout à fait pertinent de savoir si un article est cité de façon positive ou négative. Une référence bibliographique citée en contre-exemple est tout à fait révélatrice des relations entre les travaux des chercheurs. Il peut s'agir entre autre d'une référence par rapport à une définition, une hypothèse ou bien une méthode, mais également d'un point de vue, d'une comparaison ou bien d'une appréciation. Cette approche permettra également de mettre en évidence l'autocitation. La méthode de l'Exploration Contextuelle va permettre, à l'aide d'une étude poussée des indices, une analyse plus fine des références bibliographiques.

De l'importance des références bibliographiques.

Nous nous proposons d'utiliser les renvois bibliographiques d'un article afin de déterminer des segments textuels sur lesquels nous pourrons appliquer la méthode d'exploration contextuelle. L'appel

de citation dans un texte peut prendre différentes formes. Il peut s'agir principalement d'un renvoi numérique ou d'un renvoi par nom d'auteur. Pour cela, nous dresserons une classification des différentes familles numériques et alphanumériques des références bibliographiques. Afin de traiter automatiquement cette tâche d'identification et d'extraction, nous pourrons par exemple définir un alphabet adéquat permettant d'appliquer au corpus un automate fini déterministe. Cette extraction va nous permettre dans un premier temps d'étiqueter le corpus, puis de dresser des listes d'auteurs, de renvois ainsi qu'une bibliographie complète de l'auteur et de ces co-auteurs. Il sera également intéressant, dans notre approche qualitative, d'établir les relations entre les renvois bibliographiques et la bibliographie.

Approche linguistique et Exploration Contextuelle.

Suite à l'identification des appels bibliographiques, nous pourrons alors proposer une annotation de celles-ci avec une catégorie afin de définir comment l'auteur a été cité. Cette catégorisation est définie par l'étude d'indice que nous relèverons dans la phrase. Nous rechercherons les indices positifs/négatifs de citation d'un auteur, ainsi que les citations hypothèses/méthodes utilisées par un auteur. L'application des règles de l'Exploration Contextuelle permettra ainsi de lever les indéterminations sémantiques de l'unité linguistique analysée. On caractérisera ce point de vue comme étant une catégorisation sémantique des références de citation d'auteur. L'application informatique de cette étude s'effectue dans le cadre de la plateforme EXCOM (Exploration Contextuelle Multilingue) qui est en cours de réalisation au sein du Laboratoire LaLICC. Ce moteur d'annotation sémantique s'appuie sur la méthode de l'Exploration Contextuelle et permet d'étiqueter automatiquement un texte à partir de ressource linguistique. Nous serons alors en mesure d'apporter une information d'ordre sémantique et à terme de proposer une évaluation qualitative des renvois bibliographiques. Enfin cette approche proposera de dépasser le cadre bibliométrique pour analyser les sources d'un texte et détecter d'éventuelles cliques entre auteurs au sens de la théorie des graphes.

Humanities' Computings

Meurig BEYNON
Computer Science, University of Warwick, Coventry UK

Roderick R. KLEIN
Syscom Lab, University of Savoie, Bourget du Lac, France

Steve RUSS
Computer Science, University of Warwick, Coventry UK

As discussed by McCarty, Beynon and Russ in a session organised at ACH/ALLC 2005, there is a remarkable convergence between McCarty's concept of 'model-building in the role of experimental end-maker' (McCarty 2005:15) - a cornerstone in his vision for humanities computing (HC) - and the principles of Empirical Modelling (EM) (EMweb). More problematic is the tension between the pluralist conception of computing that is an essential ingredient of McCarty's stance on HC, and the prominent emphasis on 'dissolving dualities' in McCarty, Beynon and Russ (ACH/ALLC 2005:138). Resolving this tension transforms the status of HC from one amongst many varieties of computing to that of first amongst equals.

The plurality of computing In presenting his "rationale for a computing practice that is of and foras well as in the humanities", McCarty (2005:14) emphasises the plurality of computing. Following Mahoney (2005), he calls into question the search for "the essential nature of computing" and appeals to history as evidence that 'what people want computers to do' and 'how people design computers to do it' determine many different computings. The audacity of McCarty's vision in recommending his readers, as would-be practitioners of a variety of computing, to "[turn] their attention from working out principles to pursuing practice", is striking. It is hard to imagine a reputable computer science department encouraging its students to see computing primarily in terms of its practice - congenial to the students themselves as this might be. In promoting computing as an academic subject, there is no recognised

focus for developing 'scientific' principles other than the theory of computation at its historical core (Turing). McCarty instead sets out to characterise the practice of HC in such terms that it has its own integrity.

When contemplating McCarty's boldness, it is instructive to consider the alternatives. The problematic nature of the relationship between computer science and the humanities is notorious. Consider, for instance, Chesher's observation (ACH/ALLC 2005:39) that - in teaching a course in Arts Informatics: "The Humanities critiques of science and technology (Heidegger, Virilio, Coyne) are difficult to reconcile with scientific conceptions of humanities practices (Holtzman). Each of these areas places quite different, and often clearly conflicting discourses, techniques and systems of value". From this perspective, seeking to characterise HC as a unified entity seems to be the only plausible strategy, though it resembles conjuring a stable compound from an explosive combination of improbable ingredients. Invoking EM is helpful in critiquing McCarty's treatment of this unification (2005: 195-8). To elaborate the chemistry metaphor, it illuminates the precise nature of the reaction and identifies it with a more general phenomenon.

How modelling and computer science interact in humanities computing The semantic orientation of HC is crucial to understanding its chemistry. Where computer science emphasises prescribing and representing precise intended meanings, humanities is of its essence obliged to engage with meanings that are ambiguous and unintended. The authentic spirit of McCarty's notion of HC is captured in Ramsay's 'placing visualisation in a rhetorical context' (in his presentation at ACH/ALLC 2005:200) - the creative construction of an artefact as a subject for personal experience, whose interpretation is to be negotiated and potentially shared. This theme is amplified in many topical contributions to the proceedings of ACH/ALLC 20051.

Interpreting such activities from an EM perspective obliges a more prominent shift in emphasis from the accepted view of computing than is acknowledged in (McCarty, 2005) - the rich diversity of HC activities cannot be attributed primarily to the versatility of the Turing Machine as a generator of functional relationships2. In EM, the focus is upon the role that observables, dependency and agency play in the modelling activity, and each of these concepts appeals to a personal experience of interaction with technology that defies merely functional characterisation. On this basis, EM trades first and foremost not in objective 'formal' interpretations, but in speculative constructions that cannot be realised or mediated without skillful and intelligent human interaction. Appreciation of observables, dependency relationships and potential agency is acquired through developing familiarity and evolving skills. This is in keeping with Polanyi's account - cited by McCarty (2005:44) - of how awareness is transformed through skill acquisition. Functional abstraction can express transcendental computational relationships, but does not encompass such issues, which relate to what is given to the human interpreter in their immediate experience. A useful parallel may be drawn with musical performance. Though one and the same functional relationship between visual stimulus and tactile response is involved, a virtuoso pianist can perform an extended extract from a complex score in the time it takes a novice to identify the initial chord.

In this context, it is significant that - in elaborating his vision for HC, McCarty (2005:53) drew upon his experience of making a specific model - the Onomasticon for Ovid's Metamorphoses - whose construction and interpretation can be viewed as an EM archetype. Model-building in the Onomasticon, being based on spreadsheet principles, supplies the framework within which McCarty's experimental 'end-maker' role can be played out most effectively. It is implausible that the same qualities can be realised on account of adopting other model-building principles, such as the use of object-orientation, since - in conventional use - their primary purpose is to rationalise the specification of complex functional abstractions. This challenges Galey's - no doubt pragmatically most sensible! - contention (ACH/ALLC 2005:198) that "In order to bring electronic editing projects like the eNVS to the screen, humanists must think past documents to embrace the principles of object-oriented and standards-compliant programming and design.".

Humanities computing as the archetype for all varieties of computing Though McCarty (2005:14) first discusses plurality in computing in relation to communities of practice quite generally, his interest in a conceptual unification of HC and computer science (2005:195-8) acknowledges the plurality of HC itself. Where McCarty (2005:198) identifies "general-purpose

modelling software, such as a spreadsheet or database" as one component within a more diverse unity, Beynon and Russ have a radically different conceptualisation in mind. Their account identifies EM as hybrid part-automated-part-human processing within a framework for generalised computation similar to that implicit in McCarty's Onomasticon3. Within this framework, the functionality of the Turing Machine is subsumed by closely prescribed and highly automated modes of interaction, whilst modelling with the Onomasticon is a more open-ended human-centred form of processing - though by no means the most general activity of this nature. This places EM at the centre of a broader pragmatic discourse on programming that complements the conventional rational discourse (Beynon, Boyatt and Russ, 2005).

The emphasis in (McCarty, Beynon and Russ, 2005) on dissolving dualities within the frame of Radical Empiricism (James, 1996) may still appear to be mismatched to the plurality of HC. Klein's reaction to EM exemplifies the issues. In seeking a technology to support a world-wide collaborative creative venture4, he recognises the qualities of EM as supporting a concept of creativity that is expressed in the motto: "Build the camera while shooting the film" (cf. Lubart 1996, Klein 2002, METISweb). For Klein, this recognition calls to mind Joas's concept of creative action, and the processes that shape the evolving meaning of context in Lévy's 'universe in perpetual creation' (1997). The relevance of Radical Empiricism even where such diverse perspectives are being invoked stems from the subject-independent association it establishes between sense-making and the classification of relationships between experiences. For instance, whatever meaningful relationships inform the semiotics of Lévy's Information Economy Meta Language (2005) should somewhere be 'experiencable' (cf. James, 1996:160), and in this manner be amenable to EM. Seen in this light, Radical Empiricism and EM relate to universal learning activities that are orthogonal to the subject of the learning (cf. Beynon and Roe, 2004). This accords with James's monist view of experience and pluralist view of understanding (James, 1996:194). It is also resonates best with cultures where understanding through relationship has higher priority than objectification. In emphasising interaction and the interpretation of relationships, EM does not prescribe a rigid frame for understanding, but exhibits that positive quality of blandness5 (Jullien, 2004) that affords participation in many relationships. Even within the small community of EM practitioners, this potential for plurality can be seen in different nuances and idioms of elaboration, as in relation to analogue, phenomenological or ecological variants of computing.

The aspiration of EM to connect computing decisively with modelling was also that of object-oriented (OO) modelling, as first conceived nearly forty years ago (Birtwistle, Dahl, Myhrhaug and Nygaard, 1982). As a young technology, EM cannot yet compete with OO in tackling technical challenges in HC, such as devising adaptive web interfaces for the 'end-maker'. Perhaps, unlike OO, it can be more widely adopted and developed without in the process being conscripted to the cause of supporting functional abstraction. If so, it may yet demonstrate that the modelling activity McCarty has identified as characteristic of HC is in fact an integral and fundamental part of every computing practice: that all computings are humanities' computings.

Notes

1. For instance: acknowledging that there is no definitive digital representation (Galey, ACH/ALLC 2005:198); recognising the essential need for interactive playful visualisation (Ramsay, *ibid:* 200; Wolff, *ibid:* 273; Durnad and Wardrip-Fruin, *ibid:* 61); and appreciating the importance of collaborative modelling and role integration (Best et al, *ibid:* 13; van Zundert and Dalen-Oskam, *ibid:* 249).

2. For more background, see McCarty (2005) Figure 4.2 and the associated discussion on pages 195-8.

3. The framework alluded to here is that of *the Abstract Definitive Machine,* as described at (EMweb).

4. The Metis project (METISweb) is exploring collective creativity of global virtual teams of students and professionals in the movie industry.

5. The Chinese *'dan'*, which Jullien translates as *'fadeur':* Varsano notes that she "would have liked to find an English word that signifies a lack of flavor and that at the same time benefits from the positive connotations supplied by a culture that honors the presence of absence" (see Schroeder 2005).

References

ACH/ALLC 2005: *Conference Abstracts*. University of Victoria, BC, Canada, June 2005

Beynon, W.M. (2005) Radical Empiricism, Empirical Modelling and the nature of knowing, *Cognitive Technologies and the Pragmatics of Cognition*, Pragmatics and Cognition, 13:3, 615-646

Beynon, W.M. and Roe, C.P. (2004) Computer Support for Constructionism in Context, in Proc ICALT 04, Joensuu, Finland, August 2004, 216-220

Beynon W.M., Boyatt R.C., Russ S.B. (2005) Rethinking Programming, In Proc. ITNG 2006, Las Vegas, April 2006 (to appear)

Birtwistle, G M, Dahl, O-J, Myhrhaug, B, Nygaard, K. (1982) Simula Begin (2nd ed.), Studentlitteratur, Lund, Sweden, 1982

Coyne, R. (1999) *Technoromanticism, Digital narrative, holism, and the romance of the real*. Cambridge, Mass: MIT Press, 1999

Heidegger, M. (1977) *The question concerning technology, and other essays*. New York: Garland Pub.

Holtzman, S.R. (1994) *Digital Mantras. The languages of abstract and virtual worlds*. Cambridge Mass. & London: MIT Press

Joas, H. (1996) *The Creativity of Action*, Blackwell Publishers (UK)

James, W. (1996) *Essays in Radical Empiricism* (first published 1912), New York: Dover

Jullien, F. (2004) *In Praise of Blandness: Proceeding from Chinese Thought and Aesthetics*, trans. By Paula M. Varsano, Cambridge, MA: MIT Press

Klein, R.R. (2002) 'La Mètis-pour-créer ? 'Vers l'Analyse Médiologique d'une Métaphore: La Créativité Selon la lecture de l'ouvrage de François JULLIEN (1989) : Procès ou Création. Memoire de MSc Communications. Université de Nice-Sophia-Antipolis, France, 223pp

Lévy, P. (1997) *Cyberculture: rapport au Conseil de l'Europe* ed. Odile Jacob, Paris

Lévy, P. Cognitive Augmentation Languages for Collective Intelligence and Human Development. http://www.icml9.org/program/public/documents/LEVY-CRICS-205153.pdf (accessed 14/11/2005)

Lubart, T.I. (1996) *Creativity across cultures*, Handbook of creativity (R.J.Sternberg ed.), Cambridge Press.

Mahoney, M.S. (2005) The Histories of Computing(s), Interdisciplinary Science Review, 30(2), June 2005, 119-135

McCarty, W. (2005) *Humanities Computing*, Basingstoke: Palgrave Macmillan

McCarty, W, Beynon, W.M. and Russ, S.B. (2005) Human Computing: Modelling with Meaning, in ACH+ALLC 2005, 138-144

Schroeder, S. (2005) Book Review of *In Praise of Blandness: Proceeding from Chinese Thought and Aesthetics*, François Jullien. Translated by Paula M. Varsano., Essays in Philosophy, Vol. 6, No. 1, January 2005

Turing, A.M. (1936) On computable real numbers, with an application to the Entscheidung problems, Proc. London Math. Soc., 42, 230-265.

Virilio, P. (1991) *Lost Dimension*, New York: Semiotext(e) (Autonomedia)

EMweb. http://www.dcs.warwick.ac.uk/modelling/ (accessed 14/11/2005)

METISweb. http://www.metis-global.org (accessed 14/11/2005)

Using Software Modeling Techniques to Design Document and Metadata Structures ♣

Alejandro BIA
U. Miguel Hernández (Spain)

Jaime GÓMEZ
U. Alicante (Spain)

This paper discusses the applicability of modelling methods originally meant for business applications, on the design of the complex markup vocabularies used for XML Web-content production.

We are working on integrating these technologies to create a dynamic and interactive environment for the design of document markup schemes.

This paper focuses on the analysis, design and maintenance of XML vocabularies based on UML. It considers the automatic generation of Schemas from a visual UML model of the markup vocabulary, as well as the generation of DTDs and also pieces of software, like input forms.

INTRODUCTION

Most authors that treated the relationship between UML and XML [5, 7] only targeted business applications and did not consider complex document modelling for massive and systematic production of XML contents for the Web. In a Web publishing project, we need to produce hundreds of XML documents for Web publication.

Digital Library XML documents that model the structure of literary texts and include bibliographic information (metadata), plus processing and formatting instructions, are by far much more complex than the XML data we usually find in business applications. Figure 1 shows a small document model based on the TEI. Although it may seem complex, it is only a very small TEI subset.

This type of markup is not as simple and homogeneous as conventional structured data. In these documents we usually find a wide variety of elements nested up to deep levels, and there are many exceptional cases that can lead to unexpected markup situations that also need to be covered. Complex markup schemes like TEI [9] and DocBook [1] are good examples of this versatility.

However, no matter how heterogeneous and unpredictable the nature of humanities markup could get to be, software engineers have to deal with it in a systematic way, so that automatic processes can be applied to these texts in order to produce useful output for Web publishing, indexing and searching, pretty printing, and other end user facilities and services. There is also a need to reduce content production times and costs by automating and systematizing content production. For these, software, documentation and guides of good practice have to be developed.

The building of all these automation, methods and procedures within the complexity of humanities content structuring can be called Document Engineering. The purpose is to reduce costs, and facilitate content production by setting constraints, rules, methods and implementing automation wherever and whenever is possible.

XML, DTD or Schemas, XSL transforms, CSS stylesheets and Java programming are the usual tools to enforce the rules, constraints and transformations necessary to turn the document structuring problem to a systematic automated process that lead to useful Web services. But the wide variety of Schema types, and the individual limitations of each of them, make the task of setting a production environment like this very difficult.

On one hand we need a markup vocabulary that can cover all document structuring requirements, even the most unusual and complex, but that is simple enough for our purposes. In other words, we need the simplest DTD/Schema that fits our needs. We previously treated the problem of DTD/Schema simplification in [2, 3].

But DTD/Schema simplification, although useful, doesn't solve all the problems of Document Engineering, like building transformations to obtain useful output or assigning behaviour to certain structures (like popup notes, linking, and triggering services). This kind of environments are usually built incrementally. The design information, if any, is dispersed into many pieces of software (Schemas, transformation, Java applets and servlets), or does not exist at all. A system like this includes document design (DTD/Schemas), document production

techniques and tools (XSL and Java), document exploitation tools (indexing, searching, metadata, dictionaries, concordances, etc.) and Web design altogether.

UML modelling may be the answer to join all those bits and pieces into a coherent design that reduces design cost, improves the quality of the result, provides documentation and finally may even simplifies maintenance. UML modelling for massive Web content production may also lead to automatic generation of some of the tools mentioned.

ADVANTAGES OF MODELING XML DOCUMENTS WITH UML

Apart from modelling the structure of a class of documents (as DTDs and Schemas do), UML can capture other properties of elements:

- Behaviour: this is related to event oriented functions (e.g. popup notes)
- Additional powerful validation features (e.g. validating

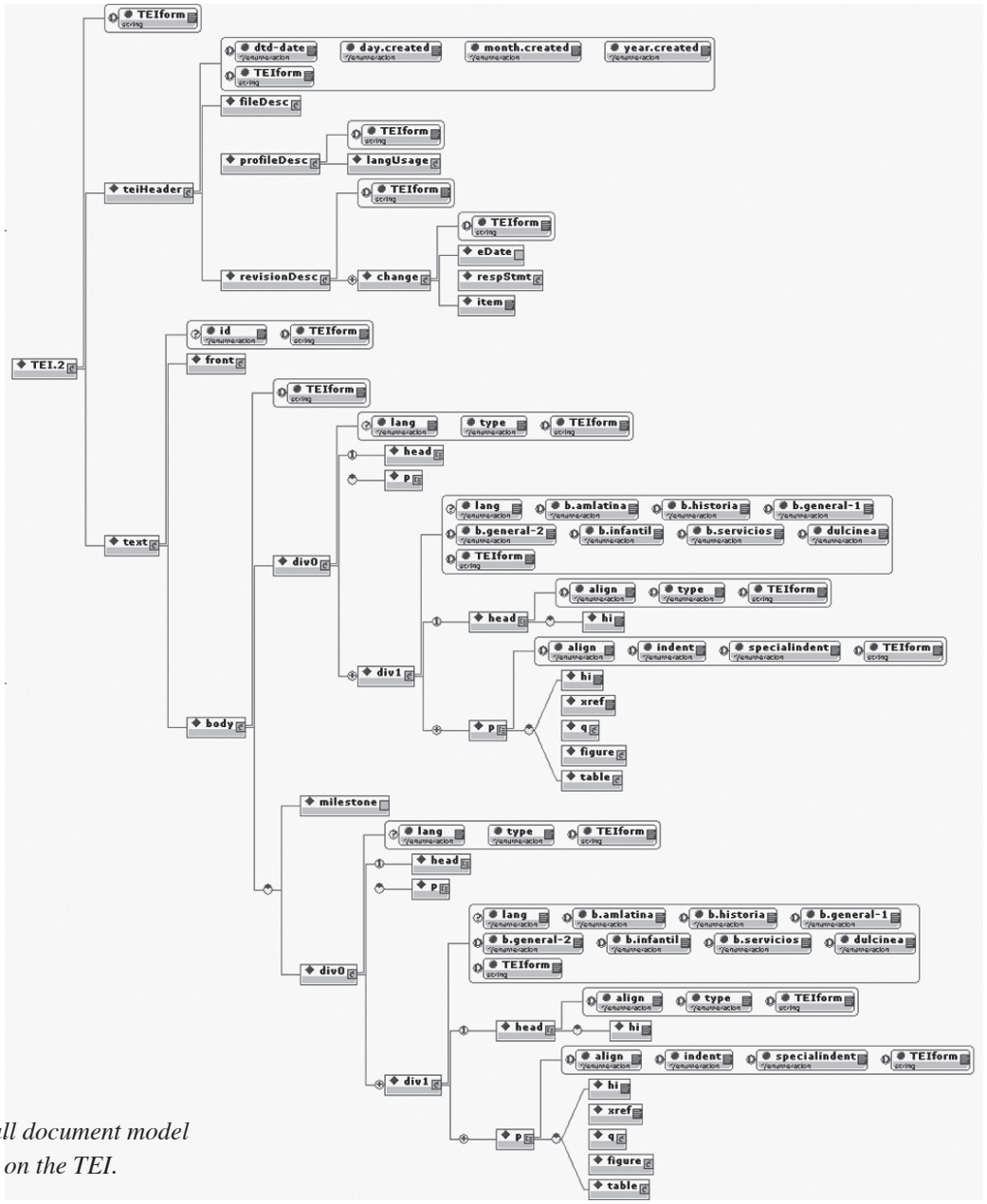


Fig. 1: a small document model based on the TEI.

consistency of certain fields like author name against a database.)

- Customization of document models to provide different views or subsets of the markup scheme to different users (e.g. DTDs for development of different types of news)

We believe that the dynamic and interactive environment described here will be very useful to professionals responsible for designing and implementing markup schemes for Web documents and metadata.

Although XML standards for text markup (like TEI and DocBook) and metadata markup (e.g. MODS, EAD) are readily available [8], tools and techniques for automating the process of customizing DTD/Schemas and addingpostprocessing functionality are not.

PREVIOUS RELATED WORK

As Kimber and Heintz define it [7], the problem is how do we integrate traditional system engineering modelling practice with traditional SGML and XML document analysis and modelling?

According to David Carlson [5], eXtensible Markup Language (XML) and Unified Modelling Language (UML) are two of the most significant advances from the fields of Web application development and object-oriented modelling.

DESCRIPTION OF THE PROJECT

We are working on integrating these technologies to create a dynamic and interactive environment for the design of document markup schemes (see figure2). Our approach is to expand the capabilities of Visual Wade[1] to obtain a tool that allows the visual analysis, design and maintenance of XML vocabularies based on UML. Among the targets we are working on the automatic generation of different types of DTD/Schemas from a visual UML model of the markup vocabulary, code generation when possible (like generating HTML forms or XSLT), documentation and special enhanced validators that can perform verifications beyond those allowed by DTDs or Schemas (like verification of certain element content or attribute values against a database).

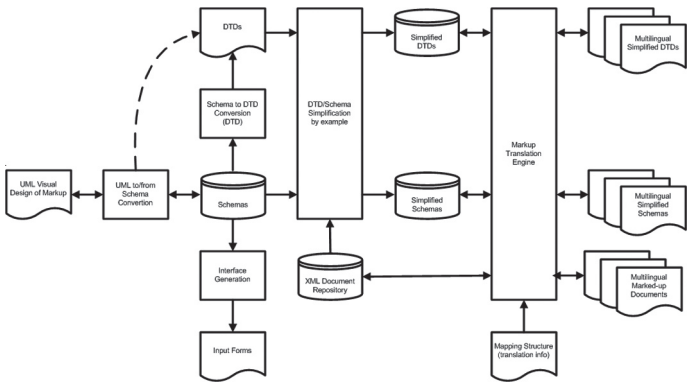


Fig. 2.: a environment for the design of document markup schemes.

Carlson [5] suggests a method based on UML class diagrams and use case analysis for business applications which we adapted for modelling document markup vocabularies.

A UML class diagram can be constructed to visually represent the elements, relationships, and constraints of an XML vocabulary (see figure 3 for a simplified example). Then all types of Schemas can be generated from the UML diagrams by means of simple XSLT transformations applied to the corresponding XMI representation of the UML model.

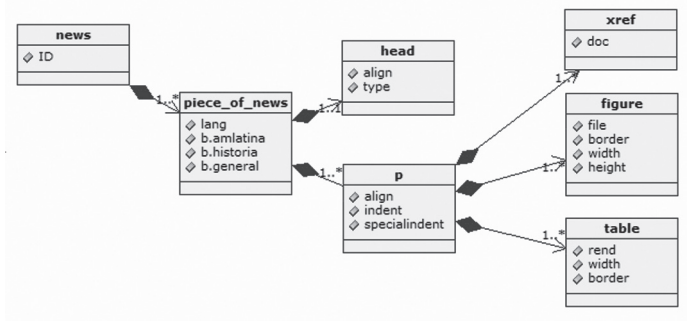


Fig. 3.: Example of a UML class diagram (partial view).

The UML model information can be stored in an XML document according to the XMI (XML Metadata Interchange) standard as described by Hayashi and Hatton [6]: "Adherence to the [XMI] standard allows other groups to easily use our modelling work and because the format

is XML, we can derive a number of other useful documents using standard XSL transformations". In our case, these documents are Schemas of various types as well as DTDs. Like Schemas, DTDs can be also generated from the XMI representation of the UML model (doted line), but as DTDs are simpler than Schemas, and all types of Schemas contain at least the same information as a DTD, DTDs can also be directly generated from them.

POSTPROCESSING AND PRESENTATIONAL ISSUES

In many cases, code generation from a high level model is also possible. Code generation may include JavaScript code to implement behaviour for certain elements like popup notes, hyperlinks, image display controls, etc. This is the case of input HTML forms that can be generated from Schemas as shown by Suleman [10].

We have successfully experimented on the generation of XSLT skeletons for XML transformation which save a lot of time. Usually XSL transforms produce fairly static output, like nicely formatted HTML with tables of contents and hyperlinks, but not much more. In exceptional cases we can find examples of more sophisticated interaction.

This high level of flexible interactivity is the real payoff from the UML-XML-XSLT-browser chain.

This sort of functionality is usually programmed specifically for individual projects, given that it's highly dependent on the nature of the markup in any given document. We aim to provide the ability to specify this at the UML level. For instance, a note could be processed differently according to its type attribute and then be displayed as a footnote, a margin note, a popup note, etc. In certain cases it can be hooked to a JavaScript function to be popped up in a message window or in a new browser instance according to attribute values. In this sense, we could provide a set of generic JavaScript functions which could retrieve content from elements and display it in various ways (popup, insertions, etc.) or trigger events (like a dictionary lookup).

We should look for document models that allow al kinds of presentation, navigation and cognitive metaphors.

- Sequential reading
- Text reuse (links and includes)
- Non-sequential reading
- Hyperlinks
- Collapsible text
- Foot notes, margin notes, popup notes
- The folder metaphor
- TOCs, indexes and menus

All the elements in a structured document have an associated semantic and a behaviour or function (as in the above example, a popup note must appear on a popup window when a link to it is pressed). This is not reflected in conventional document models: a DTD/Schema may say that a note is a popup note: ... but the behaviour of this note is not stated at all. Some postprocessing must be implemented for the popup effect to happen. A UML based document model can incorporate the expected behaviour like methods in a class diagram.

OTHER AUXILIARY TOOLS FOR DOCUMENT DESIGN AND OPTIMIZATION

As additional aiding tools for this project we have incorporated two of our earlier developments:

First the automatic simplification of DTDs based on sample sets of files [2, 3]. This tool can be applied to obtain simplified DTDs and Schemas customized to fit exactly a collection of documents.

Second, automatic element names and attribute names translation can be applied when multilingual markup is required. A detailed explanation of the multilingual markup project can be found in [4].

See figure 2 for an idea of how these tools interact with the UML document modelling.

The techniques described here can also be used for modelling metadata markup vocabularies.

CONCLUSIONS

Concerning the described set of DTD/Schema design tools, the integration of UML design with example based automatic simplification and multilingual vocabulary capabilities, is expected to be a very useful and practical design aid. However, we experienced some limitations in the use of UML. While commercial

non UML products like XML Spy or TurboXML use custom graphical tree representation to handle XML schemas, comprising very handy collapsing and navigating capabilities, most general purpose UML design environments lack these specialized features.

One of the downsides of UML is that it is less friendly when working with the low-level aspects of modelling [11]. For instance, it is easy to order the elements of a sequence in a tree, but it is very tricky to do so in UML.

Although UML proves very useful for modelling document structures of small to medium complexity (metadata applications and simple documents), UML models for medium to big sized schemas (100 to 400 elements), like those used for complex DL documents, become practically unmanageable[2]. The diagrams become overloaded with too many class boxes and lines, which end up being unreadable. This problem could be solved, or at least mitigated, by enhancing the interfaces of UML design programs with newer and more powerful display functions. Facilities like intelligent collapsing or hiding of diagram parts or elements, overview maps (see figure 3), zooming, 3-D layouts, partial views, and other browsing capabilities would certainly help to solve the problem.

Footnotes

♣ This work is part of the METASIGN project, and has been supported by the Ministry of Education and Science of Spain through the grant number: TIN2004-00779.

1 VisualWade is a tool for software development based on UML and extensions. It was developed by our research group, named IWAD (Ingeniería Web y Almacenes de Datos - Web Engineering and Data-Warehousing), at the University of Alicante. This group also developed the OOH Method (for more information see http://www.visualwade.com/)

2 The DTD used by the Miguel de Cervantes DL for its literary documents contains 139 different elements. The “teixlite” DTD, a simple and widely used XML-TEI DTD, contains 144 elements.

References

[1] **Allen, T., Maler, E., and Walsh, N.** (1997) *DocBook DTD*. Copyright 1992-1997 HaL Computer Systems, Inc., O’Reilly & Associates, Inc., Fujitsu Software Corporation, and ArborText, Inc.http://www.ora.com/davenport/

[2] **Bia, A., Carrasco, R.** (2001), Automatic *DTD Simplification by Examples*. In ACH/ALLC 2001. The Association for Computers and the Humanities, The Association for Literary and Linguistic Computing, The 2001 Joint International Conference, pages 7-9, New York University, New York City, 13-17 June 2001.

[3] **Bia, A., Carrasco, R., Sanchez, M.** (2002) *A Markup Simplification Model to Boost Productivity of XML Documents*. In Digital Resources for the Humanities 2002 Conference (DRH2002), pages 13-16, University of Edinburgh, George Square, Edinburgh EH8 9LD - Scotland - UK, 8-11 September 2002.

[4] **Bia, A., Sánchez, M., Déau, R.** (2003) *Multilingual Markup of Digital Library Texts Using XML, TEI and XSLT*. In XML Europe 2003 Conference and Exposition, page 53, Hilton Metropole Hotel, London, 5-8 May 2003. IDEAlliance, 100 Daingerfield Road, Alexandria, VA 22314.

[5] **Carlson, D.** (2001) *Modeling XML Applications with UML*. Object Technology Series. Addison-Wesley, 2001.

[6] **Hayashi, L., Hatton, J.** (2003) *Combining UML, XML and Relational Database Technologies. The Best of All Worlds For Robust Linguistic Databases*. In Proceedings of the IRCS Workshop on Linguistic Databases, pages 115--124, University of Pennsylvania, Philadelphia, USA, 11-13 December 2001. SIL International.

[7] **Kimber, W.E., Heintz, J.** (2000) *Using UML to Define XML Document Types*. In Extreme Markup Languages 2000, Montreal, Canada, 15-18 August 2000.

[8] **Megginson, D.** (1998) *Structuring XML Documents*. Charles Goldfarb Series. Prentice Hall, 1998.

[9] **Sperberg-McQueen, M., Burnard, L., Bauman, S., DeRose, S., and Rahtz, S.** (2001).*Text Encoding Initiative: The XML Version of the TEI Guidelines*. © 2001 TEI Consortium (TEI P4, Guidelines for Electronic Text Encoding and Interchange, XML-compatible edition) http://www.tei-c.org/P4X/

[10] **Suleman, H.** (2003) *Metadata Editing by Schema*. In Traugott Koch and Ingeborg Solvberg, editors, Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003, volume 2769, pages 82-87, Trondheim, Norway, August 2003. Springer-Verlag.

[11] **Marchal, B.** (2994) Design *XML vocabularies with UML tools*. March 31st, 2004, http://www-128.ibm.com/developerworks/xml/library/x-wxxm23/ or ftp://www6.software.ibm.com/software/developer/library/x-wxxm23.pdf

The Multilingual Markup Website ♣

Alejandro BIA
Juan MALONDA
U. Miguel Hernández (Spain)

Jaime GOMEZ
à l'U. Alicante (Spain)

INTRODUCTION

Markup is based on mnemonics (i.e. element names, attribute names and attribute values). These mnemonics have meaning, being this one of the most interesting features of markup. Human understanding of this meaning is lost when the encoder doesn't understand the language the mnemonics are based on. By "multilingual markup" we refer to the use of parallel sets of tags in various languages, and the ability to automatically switch from one to another.

We started working with multilingual markup in 2001, within the Miguel de Cervantes Digital Library. By 2003, we have built a set of tools to automate the use of multilingual vocabularies [1]. This set of tools translates both XML document instances, and XML document validators (we first implemented DTD translation, and then Schemas [2]). First we translated the TEI tagset, and most recently the Dublin Core tagset [3] to Spanish, and Catalan. Other languages were added later[1].

Now we present a Multilingual Markup Website that provides this type of translation services for public use.

PREVIOUS WORK

At the time when we started this multilingual markup initiative in 2001 there were very few similar attempts to be found [4]. Today they are still scarce [5, 6].

Concerning document content, XML provides built-in support for multilingual documents: it provides the predefined *lang* attribute to identify the language used in any part of a document. However, in spite of

allowing users to define their own tagsets, XML does not explicitly provide a mechanism for multilingual tagging.

THE MAPPING STRUCTURE

We started by defining the set of possible translations of element names, attribute names, and attribute values to a few target languages (Spanish, Catalan and French). We stored this information in an XML translation mapping document called "tagmap", whose structure in DTD syntax is the following:

```
<!ELEMENT tagmap (element)+ >
<!ELEMENT element (attr)* >
    <!ATTLIST element
        en CDATA #REQUIRED
        es CDATA #REQUIRED
        fr CDATA #REQUIRED>
<!ELEMENT attr (value)* >
    <!ATTLIST attr
        en CDATA #REQUIRED
        es CDATA #REQUIRED
        fr CDATA #REQUIRED>
<!ELEMENT value EMPTY >
    <!ATTLIST value
        en CDATA #REQUIRED
        es CDATA #REQUIRED
        fr CDATA #REQUIRED >
```

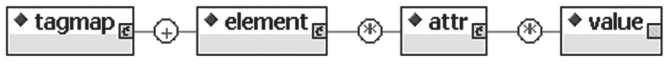


Fig. 1. Structure of the original **tagmap.xml** *file*

This structure is pretty simple, and proved useful to support the mnemonic equivalences in various languages. It was meant to solve ambiguity problems, like having two attributes of the same name in English, who should be translated to different names in a given target language. For this purpose, this structure obliges us to include all the attribute names for each element and their translations. The problem with this is global attributes, which in this approach needed to be repeated, once for each element. This made the maintenance of this file cumbersome. Sebastian Rahtz then proposed another structured, under the assumption that an attribute name has the same meaning in all cases, no mater the element it is associated to, and accordingly it would have only one target translation to a given language. This is usually the case, and although theoretically there could be cases of double meaning, as above mentioned, they do not seem to appear within the TEI. So the currently available "teinames.xml" file follows Sabastian's structure. Note that "element", "attribute" and "value" appear at the same level, instead of nested:

```
<!ELEMENT i18n (element | attribute | value)+ >
<!ELEMENT element (equiv | desc)* >
    <!ATTLIST element
        ident CDATA #REQUIRED >
<!ELEMENT attribute (equiv | desc)* >
    <!ATTLIST attribute
        ident CDATA #REQUIRED >
<!ELEMENT value (equiv)* >
    <!ATTLIST value
        ident CDATA #REQUIRED >
<!ELEMENT equiv EMPTY >
    <!ATTLIST equiv
        xml:lang CDATA #REQUIRED
        value CDATA #REQUIRED >
```

In 2004, we discussed the idea of adding brief text descriptions to each element, the same brief descriptions of the TEI documentation, but now translated to all supported languages. This would allow the structure to provide help or documentation services in several languages, as another multilingual aid. This capability was then added to the "teinames.xml" file structure, although the translations of the all the descriptions still need to be completed:

```
<!ELEMENT desc (#PCDATA) >
    <!ATTLIST desc
        xml:lang CDATA #REQUIRED >
```

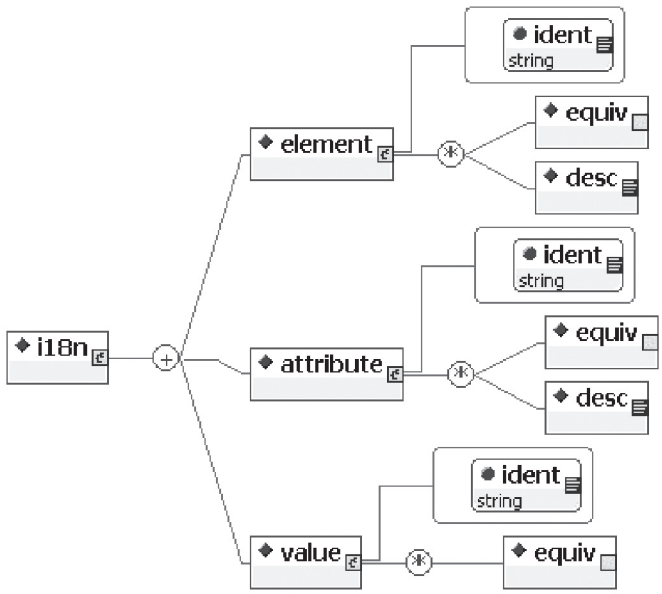


Fig. 2. Structure of the **teinames.xml** *file.*

THE MULTILINGUAL MARKUP WEB SERVICE

By means of a simple input form, the markup of a structured file can be automatically translated to the chosen target language. The user can choose a file to process (see figure 3) by means of a "Browse" button.

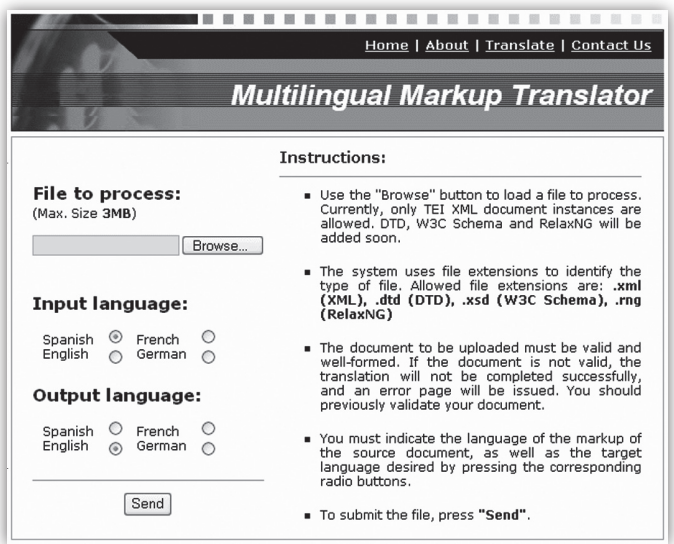


Fig. 3. The Multilingual Markup Translator form.

Currently, only TEI XML document instances are allowed. In the near future, the translation of TEI DTDs, W3C-Schemas and Relax-NG Schemas will be added, and later, other markup and metadata vocabularies will be supported, like Docbook and DublinCore.

The system uses file extensions to identify the type of file submitted. Allowed file extensions are: .xml for document instances, .dtd for DTDs, .xsd for W3C Schemas, and .rng for RelaxNG schemas.

The document to be uploaded must be valid and well-formed. If the document is not valid, the translation will not be completed successfully, and an error page will be issued. Once the source file has been chosen, the user must indicate the language of the markup of this source file, as well as the target language desired for the output. This is done by means of radio buttons.

It would not be necessary to indicate the language of the markup of the source file if it was implicit in the file itself. We thought of three ways to do this:

To use the name of the root tag to indicate the language of the vocabulary of the XML document. In this way, TEI.2 would be standard English based TEI, TEIes.2 would indicate that the document has been marked up using the Spanish tagset, and in the same way TEIfr.2, TEIde.2, TEIit.2 would indicate French, German, and Italian, for instance.

To add an attribute to the root element, to indicate the language of the tagset, for instance: <TEI.2 markupLang = "it"> would indicate that the markup is in Italian.

Use the name of the DTD to indicate the language of the tagset. TeiXLite.dtd would be English, while TeiXLiteFr.dtd would be the French equivalent.

Option 3 is by far the worst method, since a document instance may lack a DOCTYPE declaration, and there may be lots of customized TEI DTDs everywhere with very different and unpredictable names. However, options 1 and 2 are reasonably good methods to identify the language of the markup. Consensus is needed to make one of them the common practice.

IMPLEMENTATION DETAILS

For the website pages we used JSP (dynamic pages) and HTML (static pages), and these are run under

a Tomcat 5.5 web server. For the translations, we used XSLT, as described in [1, 2, 3]

AUTOMATIC GENERATION OF MARKUP TRANSLATORS USING XSLT

The XSLT model is thought to transform one input XML file into one output file (see figure 4), which could be XML, HTML, XHTML or plain text, and this includes program code. It does not allow the simultaneous processing of two input files.

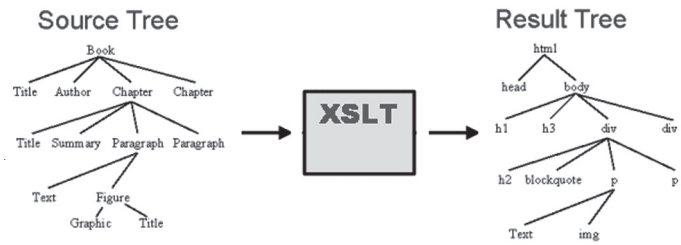


Fig. 4. The XSLT processing model.

There are certain cases when we would like to process two input files altogether, like markup translation (see figure 5).

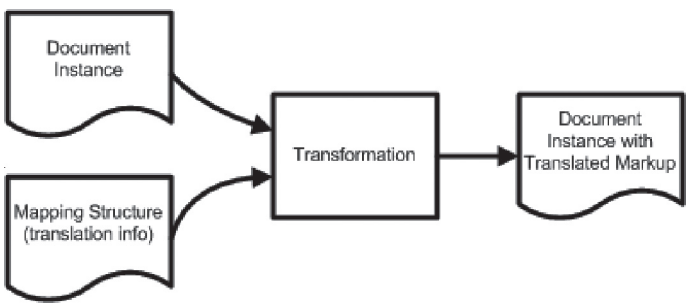


Fig. 5. The ideal transformation required.

As XSLT does not allow this, two alternatives occurred to us, both comprising two transformation steps.

The first approach is to automatically generate translators. As Douglas Schmidt said: "I prefer to write code that writes code, than to write code" [7]. This is what we have done for the MMWebsite, i.e. to pre-process the translation map in order to generate an XSLT translation script which includes the translation knowledge embedded in its logic. Then this generated script can perform all the document-instance translations required. The mapping structure supports the language equivalences for various languages, so we should generate a translator for every possible pair of languages. Whenever the mapping structure is modified, a new set of translators must be generated. Fortunately, this is an automated process.

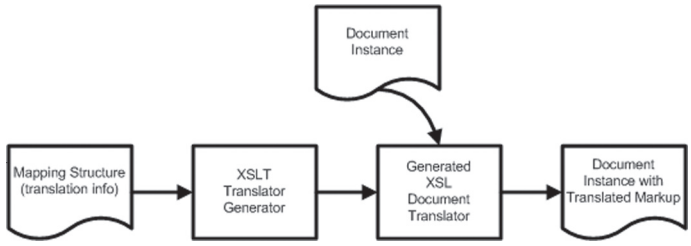


Fig. 6. Pre-generation of a translating XSLT script, to then translate the document instance.

The other alternative would be to merge the two input files into a new single XML structure, and then to process such file which would contain both the XML document instance, and the translation mapping information (see figure 7). This implies joining the two XML tree structures as branches of a higher level root.

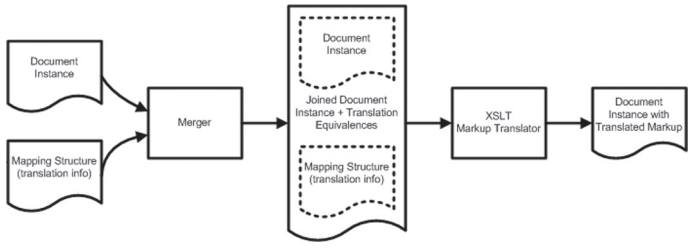


Fig. 7. Merging the two files before applying XSLT.

Although this approach may prove useful for some problems, we did not use it for the MMWebsite, because the file merging preprocessing must be done for each file to translate, increasing the web service response time. Using preprocessed translators instead proved to be a faster solution.

This limitation, which is proper of the XSLT processing model, could be avoided by using a standard programming language like Java instead.

HOW WE ACTUALLY DO IT

The mapping document which contains all the necessary structural information to develop the

language converters is read by the transformations generator, which was built as an XSLT script. XSL can be used to process XML documents in order to produce other XML documents or a plain text document. As XSL stylesheets are XML, they can be generated as an XSL output. We used this feature to automatically generate both an English-to-local-language XSL transformation and a local-language to English XSL transformation for each of the languages contained in the multilingual translation mapping file. In this way we assured both ways convertibility for XML documents (see figure 8).

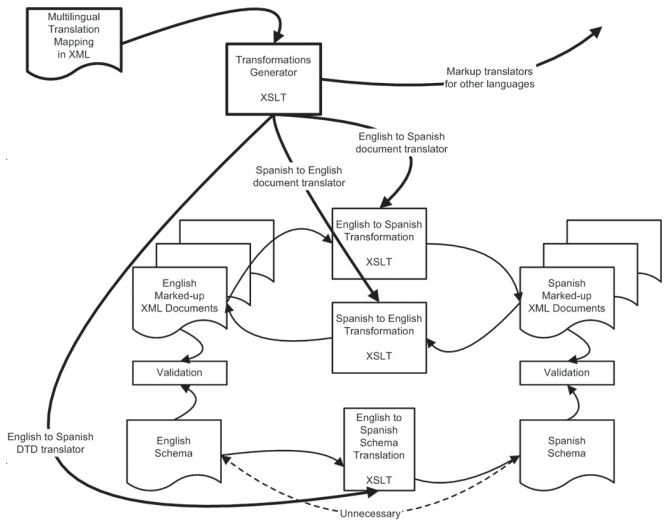


Fig. 8. Schema translation using XSLT.

For each target language we also generate a DTD or a Schema translator. In our first attempts, this took the form of a C++ and Lex parser. Later, we changed the approach. Now we first convert the DTD to a W3C Schema, then we translate the Schema to the local language, and finally we can (optionally) generate an equivalent translated DTD. This approach has the advantage of not using complex parsers (only XSLT) and also solves the translation of Schemas. In our latest implementation, the user can freely choose amongst DTD, W3C Schema and RelaxNG, both for input and output, allowing for a format conversion during the translation process.

Many other markup translators can be built to other languages in the way described here.

CONCLUSIONS

Amongst the observed advantages of using markup in one's own language are: reduced learning times, reduction of errors and higher production. It may also help spread the use of XML vocabularies like DC, TEI, DocBook, and many others, into non-English speaking countries. Cooperative multilingual projects may benefit from the possibility of easily translating the markup to each encoder's language. Last, but not least, scholars of a given language feel more comfortable tagging their texts with mnemonics based on their own language.

FUTURE WORK

Multilingual Help Services: As already said, brief descriptions for elements and attributes in different languages have been added to the mapping structure. This allows for multilingual help services, like generating a glossary in the chosen language of the elements and attributes used in a given document, or a given DTD/Schema. We are working on adding this feature.

Footnotes

♣ This work is part of the METASIGN project, and has been supported by the Ministry of Education and Science of Spain through the grant number: TIN2004-00779.

1 Translations of the TEI tagset by: Alex Bia and Manuel Sánchez (Spanish), Régis Déau (French), Francesca Mari (Catalan), Arno Mittelbach (German)

References

Endnotes

[1] **Bia, A., Sánchez, M., and Déau, R.** (2003) *Multilingual Markup of Digital Library Texts Using XML, TEI and XSLT*. In XML Europe 2003 Conference and Exposition, Organized by IDEAlliance, 5-8 May 2003, Hilton Metropole Hotel, London, p. 53, http://www.xmleurope.com/

[2] **Bia, A., and Sanchez, M.** (2004) *The Future of Markup is Multilingual*. In ACH/ALLC 2004:

Computing and Multilingual, Multicultural Heritage. The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, 11-16 June 2004, Göteborg University, Sweden, p 15-18, http://www.hum.gu.se/allcach2004/AP/html/prop119.html

[3] **Bia, A., Malonda, J., and Gómez, J.** (2005) *Automating Multilingual Metadata Vocabularies*. In DC-2005: Vocabularies in Practice, Eva Mª Méndez Rodríguez (ed.), p. 221-229, 12-15 September 2005, Carlos III University, Madrid. ISBN 84-89315-44-2. http://dc2005.uc3m.es/

[4] **Pei-Chi WU** (2000) *Translation of Multilingual Markup in XML*. In International Conference on the theories and practices of Electronic Commerce, Part II, Session 14, pages 21-36, Association of Taiwan Electronic Commerce, Taipei, Taiwan, October 2000. http://www.atec.org.tw/ec2000/PDF/14.2.pdf

[5] **Bryan, J.** (2002) *KR's Multilingual Markup*, TechNews Volume 8, Number 1: January/February 2002 http://www.naa.org/technews/TNArtPage.cfm?AID=3880

[6] **Cover, R.** *Markup and Multilingualism*, last visited online 2005-4-25 at Cover Pages: http://xml.coverpages.org/multilingual.html

[7] **Schmidt, D.** (2005) *Opening Keynote*, MoDELS 2005: ACM/IEEE 8th International Conference on Model Driven Engineering Languages and Systems, Montego Bay, Jamaica, 2-7 October 2005.

Le résumé automatique dans la plate-forme EXCOM

Antoine BLAIS
Jean-Pierre DESCLÉS
Brahim DJIOUA

Laboratoire LaLICC UMR 8139,
Paris-Sorbonne (Paris IV)

Automatic summarization appears to be for the future an important domain in textual mining and information retrieval. The main purpose of automatic summarization is the extraction of the relevant information contained in a document. There are various approaches to make a summary, some approaches use a semantic representation of the text to generate a summary, and others find relevant parts of a text to extract them and to constitute the summary. In the first place, we introduce the domain of automatic summarization and then we present the two main approaches. Finally, we expose our method, ours choices, and the software application which proceed from it.

1. Présentation du résumé automatique

Le résumé automatique a pour but de fournir à un utilisateur l'information pertinente et essentielle d'un document sous forme de rapport synthétique. Il doit au travers d'un résumé retranscrire le sens général de ce que le document original à voulu exprimer. Le parcours d'un document pour connaître son intérêt lors d'une recherche par un utilisateur peut être long et inutile, notamment s'il doit parcourir un grand nombre de textes. Un logiciel de résumé automatique permet ainsi à des utilisateurs de ne pas parcourir les textes dans leur totalité, en ne leur faisant lire que le résumé, ce qui produit ainsi pour eux un gain de temps important dans leurs recherches. L'intérêt du résumé automatique apparaît pour la consultation sélective de documents par des personnes dont la lecture entière de ceux-ci est impossible. Permettant une lecture synthétique, il aide le lecteur à filtrer les documents qui peuvent l'intéresser pour une lecture ultérieure alors normale du ou des documents choisis selon sa recherche.

On note que la nature du résumé reste floue, il n'existe pas en effet de résumé standard. De nombreuses expériences sur des résumeurs professionnels montrent qu'il n'existe pas de résumé type.

Le laboratoire *LaLICC* de l'université Paris4-Sorbonne a par ailleurs acquis une expérience dans le domaine du résumé automatique depuis plusieurs années. La réalisation de plusieurs projets tels que SERAPHIN, SAFIR et ContextO ont participé à la réflexion et à la mise en place d'applications concrètes dans ce domaine. Dans la phase actuelle, ces travaux entrepris sont repris en étant introduit dans la nouvelle plate-forme informatique EXCOM (EXploration COntextuelle Multilingue) qui a pour objectif principal l'annotation sémantique automatique de texte (dont la tâche de résumé automatique fait partie).

2. Les différentes approches du résumé automatique

Nous allons présenter ici les deux grandes méthodes existantes dans le résumé automatiques afin d'introduire ensuite le projet EXCOM.

La méthode par compréhension

Cette méthode est issue essentiellement du domaine de l'intelligence artificielle. Elle considère la tâche de résumé automatique comme devant être calquée par l'activité résumante humaine. La constitution d'un résumé par un logiciel doit ainsi passer par la compréhension totale du texte. Le logiciel doit pouvoir construire une représentation du texte, qui éventuellement peut être modifier ensuite, afin de pouvoir générer à partir de celle-ci un résumé. L'avantage de cette méthode est de vouloir s'inspirer des processus cognitifs humains utilisés dans la compréhension de texte. Néanmoins en dehors de cet aspect, des problèmes surgissent. Premièrement, la compréhension de texte par l'homme est une tâche très loin d'être comprise, donc son implémentation informatique semble encore impossible. Deuxièmement, la représentation d'un texte est également très compliquée, et cette notion reste encore difficile pour les linguistes. Chaque méthode par compréhension propose une représentation propre, mais aucune n'arrive à représenter le texte correctement. La complexité d'un texte sous tous ses aspects (discursif, temporel, etc.) est toujours une barrière à la construction correcte d'une représentation. Enfin la génération du résumé qui apparaît comme étant l'étape finale est aussi difficile. Les travaux sur la production automatique de textes à partir de représentations sont encore très limitées dans leurs résultats.

La méthode par extraction

Cette méthode est issue essentiellement du domaine de la recherche d'information. L'objectif de cette méthode est de fournir rapidement un résumé simple à valeur informative pour l'utilisateur. Elle consiste par l'extraction des phrases les plus pertinentes du texte traité afin de constituer le résumé devant retransmettre l'essentielle de l'information pertinente générale qui se dégage du texte original. Le résumé est alors constitué des phrases extraites du document. Le travail principal se situe alors dans l'évaluation de la pertinence des phrases du texte suivant un ou plusieurs critères. On peut dissocier alors deux grandes façons de faire. Les techniques statistiques qui prennent comme critère de pertinence la présence de termes fortement représentatifs du texte traité (grâce à un calcul de fréquence). Une phrase est alors extraite ou non suivant la présence de ces termes représentatifs dans celle-ci. Ces techniques sont limitées et se trouvent confrontées à certains problèmes, comme la synonymie des termes par exemple. Les techniques plutôt linguistiques s'appuient sur la présence de marques linguistiques de surfaces pour établir l'importance ou non d'une phrase dans le texte. Certaines marques bien précises permettent d'attribuer une valeur sémantico-discursive à la phrase et ainsi de connaître sa pertinence ou non dans la structure discursive du texte. L'avantage de la méthode par extraction est de ne pas passer par des représentations complexes du texte, et de pouvoir fournir un résumé de façon assez simple (en comparaison d'une méthode par compréhension). Néanmoins les problèmes surviennent dans la qualité du résumé obtenu. Comme le résumé est le résultat de l'extraction d'un ensemble de phrases du texte que l'on a concaténée, la cohésion et la cohérence du résumé peuvent devenir médiocre. Il faut donc dans ces méthodes veiller à la qualité du résumé en sortie, notamment par des méthodes d'évaluations.

3. La plate forme EXCOM et le résumé automatique

La plate-forme EXCOM est un moteur d'annotation sémantique travaillant à partir de ressources linguistiques préalablement rentrées par des linguistes.

D'un point de vue technique EXCOM repose essentiellement sur les technologies XML. Par ailleurs la plate-forme propose une ouverture vers le multilinguisme en prenant en compte d'autres langues que le français tels que l'arabe et le coréen. La technique utilisée pour l'annotation est celle de la méthode d'exploration contextuelle constituée au sein du laboratoire. Cette méthode recherche à identifier des indicateurs linguistiques dans le texte, puis dans le cas où ils seraient présents, explorer le contexte textuel dans lequel ils se situent à la recherche d'autres indices linguistiques afin de pouvoir attribuer une annotation sémantique sur le segment textuel désigné par le linguiste. Ce traitement textuel repose sur deux hypothèses fondamentales : la première admet la présence dans un texte de marques discursives affectées à des points de vue, et la seconde affirmant l'invariance de ces points de vue suivant les domaines traités dans le document. Le choix de points de vue adaptés est ainsi en rapport avec la nature du texte traitée : articles scientifiques, articles de journaux, essais, etc… L'essentiel des ressources utilisées correspond donc à un ensemble de règles d'explorations préalablement construites par les linguistes. Il convient de remarquer que cette méthode d'exploration contextuelle ne fait pas appel à des ontologies externes mais que le système reste entièrement compatible avec celles-ci.

La tâche de résumé automatique, actuellement en développement sous EXCOM, utilise pour la constitution de résumé une méthode par extraction de phrases basée sur quatre critères. A la suite de la segmentation préalable du texte en phrases, on attribue à chaque phrase quatre valeurs, correspondant aux quatre critères de pertinence.

Le premier critère de pertinence que nous avons retenu pour une phrase est la valeur de son annotation sémantique qui est attribuée par une règle d'exploration contextuelle. Les principaux points de vue que nous retenons pour le résumé sont l'annonce thématique, la conclusion, la récapitulation et les soulignements de l'auteur.

Le second critère correspond à la position de la phrase dans la structure textuelle. La position de certains types de phrases (comme les annonces thématiques ou les conclusions) par rapport à l'organisation des éléments constitutifs de l'argumentation de l'auteur, est déjà une information essentielle pour l'attribution du rôle de la phrase et de sa pertinence au niveau discursif. Ce second critère se trouve ainsi fortement lié au premier.

Le troisième critère est lié à la thématique présente dans le texte. Nous cherchons dans les phrases des *termes de filtrage*, c'est-à-dire qu'ils correspondent aux mots les plus représentatifs de l'univers thématique qui se trouve dans le texte.

Enfin le dernier critère, qui est un critère négatif, est la présence ou non dans la phrase d'anaphores pronominales. La présence de pronoms personnels sans référent dans le résumé contribue à sa mauvaise lisibilité, et des stratégies discursives devront alors être étudiées pour la sélection ou non de ces phrases.

Il existe donc trois étapes fondamentales dans la construction du résumé :

- la première étant la phase d'annotation du texte selon les points de vue
- la seconde étant la construction du résumé par la sélection des phrases disposant de la meilleure valeur de pertinence P. La valeur de pertinence P d'une phrase correspond à une valeur numérique qui est calculée en fonction des quatre critères qualitatifs qui sont affectés à chaque phrase
- enfin la troisième étape étant la phase de nettoyage du résumé obtenu dans la seconde étape à l'aide de règles appropriées, afin d'assurer une cohésion et une cohérence meilleure.

Nous montrerons donc dans la présentation des exemples de résumés que nous commenterons en expliquant l'avantage de notre stratégie.

Third-Party Annotations in the Digital Edition Using EDITOR

Peter BOOT

Huygens Instituut, Department of e-Research

Recent discussion about the scholarly digital edition has focussed on ways to change the edition from a passive text, only there to be consulted, into a dynamic research environment. Siemens in (Siemens 2005) asks why as yet we have seen no convincing integration between text analysis tools and the hypertext edition. Best in (Best 2005) speculates on the possibilities this vision offers for Shakespeare research. To some extent it seems to be what Mueller is realising in the Nameless Shakespeare (Mueller 2005). An essential step towards seamless integration of text analysis tools into the digital edition (TAML, the Text Analysis Mark-up Language) is suggested in (Sinclair 2005).

The most visionary statement of the dynamic edition's potential is no doubt given by Robinson in (Robinson 2003). A dynamic edition, in his view, while offering text analysis and other tools that may shed light on some aspect or another of the edited texts, would also be open to the addition of new content, of corrections, of many different types of annotations.

Integrating third-party annotations into the edition is something that seems especially interesting, as it would open up the edition to the results of interpretive studies. The output of scholarly processes of textual analysis (as e.g. suggested in Bradley 2003) could be fed back into the digital edition, and made available for querying by other scholars.

This paper will focus on a solution for adding third party annotations into a digital edition. It will propose a REST (Representational State Transfer, Fielding 2000) API for the exchange of annotation information between an edition and an annotation server. The edition display software (which transforms the edition XML source file into HTML) will ask the annotation server for the annotations that apply to text fragments that are being displayed to the edition user. Depending on the parameters the annotation server will return either annotations formatted for display or instructions for hyperlinking the text to the annotations. Thus, the digital edition will be able to include a display of external a nnotations without knowing about the annotations' contents or even the annotation data model.

The paper presentation will include a brief demonstration of a prototype implementation of the protocol. The demonstration will be based on a digital emblem book edition at the Emblem Project Utrecht (http://emblems.let.uu.nl) and use the EDITOR annotation toolset under development at the Huygens Institute (http://www.huygensinstituut.knaw.nl/projects/editor, Boot 2005). EDITOR at present consists of an annotation input component that runs on the user's workstation and an annotation display component that runs on a web server. The input component displays the CSS-styled edition XML to the user and facilitates the creation of multi-field user-typed annotations to arbitrary ranges of text in the edition. The display component, still at an early stage of development, shows the annotations in conjunction with the edition XML, has some facilities for filtering and sorting, and will offer, one day, advanced visualisation facilities. The EDITOR server component will serve up the annotations for display in other contexts, first and foremost, presumably, in the context of the digital edition that they annotate.

As Robinson notes, one of the more complex issues in annotating the digital edition is the problem of concurrent hierarchies and the mark-up overlap problems to which this gives rise. The EDITOR annotation toolset assumes the edition and its annotations will be stored in separate locations. Each annotation stores information about the start and end locations of the text fragment to which it applies. There is no need to materialize the annotations into tagging interspersed between the basic edition mark-up, and the overlap issue therefore does not arise (the solution in that respect is similar to the Just In Time Markup described in Eggert 2005). Similarly, as the edition XML remains unmodified, there is no need to worry about potential corruptions during the annotation process.

Making available third-party annotations from within the digital edition will go a long way towards establishing a 'distributed edition fashioned collaboratively', to borrow Robinson's words. My paper will briefly look at some of the wider issues the integration of third-party scholarship into the digital edition raises. How will the presence of

third-party material influence the edition's status? Should there be a review process for third-party contributions? Or is it old-fashioned to even think in terms of 'third parties'? Robinson speaks of the edition as a 'mutual enterprise'. Editorial institutes, such as the Huygens Institute, will need to rethink their role, as scholarly editions evolve into centrepieces of ever-expanding repositories of text-related scholarship.

References

Best, Michael (2005), 'Is this a vision? is this a dream?': Finding New Dimensions in Shakespeare's Texts', *CH Working Papers*, http://www.chass.utoronto.ca/epc/chwp/Casta02/Best_casta02.htm, accessed 2005-11-13.

Boot, Peter (2005), 'Advancing digital scholarship using EDITOR', *Humanities, Computers and Cultural Heritage. Proceedings of the XVI international conference of the Association for History and Computing 14-17 September 2005* (Amsterdam: Royal Netherlands Academy of Arts and Sciences).

Bradley, John (2003), 'Finding a Middle Ground between 'Determinism' and 'Aesthetic Indeterminacy': a Model for Text Analysis Tools', *Lit Linguist Computing,* 18 (2), 185-207.

Eggert, Paul (2005), 'Text-encoding, Theories of the Text, and the 'Work-Site', *Lit Linguist Computing,* 20 (4), 425-35.

Fielding, Roy Thomas (2000), 'Architectural Styles and the Design of Network-based Software Architectures', (University of California).

Mueller, Martin (2005), 'The Nameless Shakespeare', *CH Working Papers*, http://www.chass.utoronto.ca/epc/chwp/Casta02/Forest_casta02.htm, accessed 2005-11-13.

Robinson, Peter (2003), 'Where we are with electronic scholarly editions, and where we want to be', *Jahrbuch für Computerphilologie*, http://computerphilologie.uni-muenchen.de/jg03/robinson.html, accessed 1005-11-13.

Siemens, Ray (with the TAPoR community) (2005), 'Text Analysis and the Dynamic Edition? A Working Paper, Briefly Articulating Some Concerns with an Algorithmic Approach to the Electronic Scholarly Edition', *CH Working Papers*, http://www.chass.utoronto.ca/epc/chwp/Casta02/Siemens_casta02.htm, accessed 2005-11-13.

Sinclair, Stéfan (2005), 'Toward Next Generation Text Analysis Tools: The Text Analysis Markup Language (TAML)', *CH Working Papers*, http://www.chass.utoronto.ca/epc/chwp/Casta02/Sinclair_casta02.htm, accessed 2005-11-13.

Orlando Abroad : Scholarship and Transformation

Susan BROWN
School of English and Theatre Studies

Patricia CLEMENTS
Isobel GRUNDY
Department of English studies
University of Guelph

The Orlando Project, which has regularly reported on its work in progress at meetings of ACH/ALLC, is due for publication in early 2006. In this paper the three originating literary scholars on the project will look back at the its original goals, consider significant turning-points in the process, and reflect on what the project ended up producing for the first release.

The project takes its name from Virginia Woolf's fantastic narrative of an aspiring English writer who begins life as an Elizabethan male and is transformed in the course of the novel's romp through history into a woman who lives up to the year of the novel's publication in1928. The transformation occurs while Orlando is abroad as ambassador extraordinary for King Charles II in Turkey. Woolf's narrator has much to say about the difficulties this poses for the historian, lamenting that "the revolution which broke out during his period of office, and the fire which followed, have so damaged or destroyed all those papers from which any trustworthy record could be drawn, that what we can give is lamentably incomplete." The charred fragments offer some clues, but "often it has been necessary to speculate, to surmise, and even to use the imagination." We would locate the electronic Orlando, like Woolf's protagonist, as the site of an extraordinary transformation associated with the challenges of moving between cultures, the limitations of paper, and the necessity for speculation, imagination, and new approaches to scholarship.

None of us were experienced in humanities computing when the project was begun; we set out to write a feminist literary history. In the process of trying to figure out how to do it, we decided to use computers. Ten years later, we have produced an extensively tagged XML textbase comprising a history of women's writing in the British Isles in a form quite unlike any previous history of writing. At a glance, it may look like another translation into electronic form of the genre of alphabetical companion or literary encyclopedia, and it is indeed deeply indebted to that flexible and enduring form. However, our custom markup system has been used to tag a wide range of semantic content in the textbase, and a backend indexing and delivery system allows that markup to be exploited to produce on-the-fly documents composed of portions of the project's digitally original source documents. The result is a very flexible and dynamic approach to literary history that challenges users to harness the power of the encoding to pursue their own interests.

Recent argument about the crisis in scholarly publishing, such as that marshaled by Jerome McGann in support of the NINES initiative, has focused on the need to draw a larger community of scholars into best-practice methods of electronic markup and publication of texts. This is both crucial as a means of addressing the crisis, and indispensable to the continued development of electronic tools to serve the humanities research community. We offer ourselves as a kind of case study in such a process, given that our project did not originate as humanities computing endeavour but was completely transformed in the course of becoming one. In going electronic, we became radically experimental, tackling problems and producing results that we could not have foreseen at the outset.

We will reflect on the results of taking an already very ambitious project electronic in relation to a range of factors including:

- the impact on the intellectual trajectory of the project of engaging with, in addition to our disciplinary subject matter, a whole new field of inquiry and undertaking what became an interdisciplinary experiment in knowledge representation;
- the impact on the project's temporal trajectory;
- funding, and its relationship to funding structures and opportunities;
- the intensification of collaboration, increase in project personnel, and transformation of roles and responsibilities;
- the impact on research and writing methods of

composing an extensive scholarly text in XML;

- o the shaping of the research itself by the use of XML;
- o the development a delivery system that aimed at once to be reassuringly accessible and to challenge users to employ the system in new ways;
- o dilemmas regarding modes of publication

While the paper will, given the constraints of time, necessarily touch briefly on some of these various areas, these reflections will be framed as an inquiry into what it means to bridge the gap between the community of researchers deeply invested in humanities computing and the wider scholarly community.

We have come to see Orlando as a kind of emissary of humanities computing, in that we hope it will prove to be a major step towards establishing methods for encoding critical scholarly materials. It provides a test case of the feasibility and benefits of employing XML to encode large semantic units of critical discourse. It offers a model which we hope will be employed and adapted by other projects, and we will indicate the direction we would like to take the project in the future. But perhaps most importantly, the Orlando Project offers a substantial resource that in its design will, we hope, alert scholars beyond the humanities computing community to the potential of encoding as a means of scholarly inquiry and a tool of critical expression. The proof of that will be in the pudding, of course, so the paper will also report as far as possible on the initial reaction from the scholarly community in the field of English studies to the project's public release.

Asynchronous Requests for Interactive Applications in the Digital Humanities

Kip CANFIELD

Information Systems, University of Maryland

Web applications are becoming more sophisticated and offer a rich user experience that is similar to native client applications. Examples of these applications include Google maps and Flickr. Interactive applications in the humanities can use some of these same design patterns to improve the user experience. The asynchronous update pattern is evaluated in this paper.

Asynchronous update allows web applications to update without a complete page reload and goes by the popular name of Ajax. "Traditional web applications essentially submit forms, completed by a user, to a web server. The web server responds by sending a new web page back. Because the server must submit a new page each time, applications run more slowly and awkwardly than their native counterparts.

Ajax applications, on the other hand, can send requests to the web server to retrieve only the data that is needed - usually using SOAP or some other XML-based web services dialect. On the client, JavaScript processes the web server response. The result is a more responsive interface, since the amount of data interchanged between the web browser and web server is vastly reduced. Web server processing time is also saved, since much of it is done on the client." (from http://en.wikipedia.org/wiki/AJAX)

This paper presents a case study of an ongoing project to create a digital library of Navajo language texts. After such texts are put into the database, the texts can be annotated with interlinear linguistic information using an interactive web application. The model for interlinear information that exists in TEI was determined to be inadequate for the present application and a different model is used. The design of the application for interactive and collaborative entry of interlinear linguistic information consists of a browser client using

JavaScript and a server-side Exist native XML database that responds to XQueries. The acquisition and parsing methods for the texts are described in Canfield 2005.

The Navajo language texts with annotated interlinear information are compliant with a Relaxng schema based on the XML from Bow 2003. Since this is a pilot application, it allows the schema to be tested and perhaps modified before finally adding the schema to TEI using an ODD specification. For example, the interlinear XML for the sentence “t’11’1ko [a’ sisi[“ is:

```
<phrase>
    <item type=”txt”>t’11’1ko [a’ sisi[</item>
    <item type=”gls”>then one grabbed me</item>
    <words>
     <word>
        <item type=”txt”>t’11’1ko</item>
        <item type=”gls”>just then</item>
        <item type=”pos”>adv</item>
     </word>
     <word>
        <item type=”txt”>[a’</item>
        <item type=”gls”>one</item>
        <item type=”pos”>pro</item>
     </word>
     <word>
        <item type=”txt”>sisi[</item>
        <item type=”gls”>3p grabbed me</item>
        <item type=”pos”>verb</item>
        <morphemes>
        <morph>
                <item type=”txt”>shi</item>
                <item type=”gls”>me</item>
                <item type=”pos”>1st person obj
                </item>
        </morph>
        <morph>
                <item type=”txt”>yii</item>
                <item type=”gls”>he/she/it</item>
                <item type=”pos”>3p subj pro</item>
        </morph>
            <morph>
                <item type=”txt”>NULL</item>
                <item type=”pos”>classifier</item>
            </morph>
            <morph>
                <item type=”txt”>zi[</item>
                <item type=”gls”>grab</item>
                <item type=”root”>ZIID(1)</item>
            </morph>
        </morphemes>
     </word>
    </words>
</phrase>
```

All the XML in the database is transformed to XHTML for the user interface of the application. This page uses a tabular interface to allow the user to see and update the interlinear information for each sentence in the text. For example, if a word has already been annotated, it will appear with all the annotated information. If the word has not been annotated, the user double clicks on the word and the word appears in an editable html text input box. The user can then edit the detailed interlinear information for the text informed by an on-line Navajo lexicon. The display for each sentence appears as below, but the font for Navajo is not active so the characters display as the base ASCII. The sentence is « t’11’1ko [a’ sisi[« which means «then one grabbed me.» All fields can be edited when double-clicked except for the base word in each sentence. Note that the last word is a verb «sisi[» and each underlying morpheme is annotated.

t’11’1ko [a’ sisi[			
t’11’1ko gls=just then pos=adverb			
[a’ gls=one pos=pronoun			
sisi[gls=3rd person grabbed me pos=verb			
shi	yii	NULL	zi[
me	he/she/it		grab
1st person object	3rd person subject pronoun	classifier	stem root=ZIID(1)
gls= then one grabbed me			

The JavaScript code that updates each of the fields uses XML HTTP Request which allows the page to be updated without a page reload. Note that traditional web applications must reload the entire page for each update. This is time consuming and disruptive to the user experience.

A sample of 25 chapter-length documents was used for this evaluation from the Navajo language digital library. The average document size was about 150kb, which is not very large for documents common in humanities applications. Each document update was timed (using JavaScript) in each mode - with asynchronous requests and with traditional page re-loads. The average time for an update of a single field using XML HTTP Request was about 8 ms. The average time for a traditional (whole page reload) update of a field was about 400 ms. The traditional method shows a large update time that will cause unneeded user waiting. The traditional method also makes for a disruptive user experience where the page visually reloads while the user is trying to accomplish a task.

Many interactive web applications in the humanities would benefit from this asynchronous update design pattern. Whenever a document is large and requires many small updates, the user experience will be improved with asynchronous requests due to shorter load times and a smoother experience with the user interface.

References

Bow, C., Hughes, B., Bird, S. (2003). *Towards a General Model of Interlinear Text*. Proceedings of EMELD Workshop 2003: Digitizing & Annotating Texts & Field Recordings. LSA Institute: Lansing MI, USA. July 11-13, 2003.

Canfield, K. (2005). *A Pilot Study for a Navajo Textbase*. Proceedings of ACH/ALLC Conference 2005. Victoria, BC, Canada, June 15-18, 2005.

Problems with Marriage: Annotating Seventeenth-Century French Engravings with TEI and SVG

Claire CARLIN
Dept. of French, University of Victoria

Eric HASWELL
HCMC, University of Victoria

Martin HOLMES
HCMC, University of Victoria

THE TEXTS

This image markup project fits into the larger context of an electronic anthology, "Le marriage sous l'Ancien Régime: Une anthologie critique." Since 1998, C. Carlin has been collecting texts about early modern marriage in France for her forthcoming book, *L'imaginaire nuptial en France, 1545-1715*. Given that the majority of documents studied for the book have not been republished since their original appearance in the sixteenth and seventeenth centuries, the idea of an electronic anthology should be appealing to scholars in several disciplines (history, literary studies, linguistics, cultural studies, art history, philosophy, religious studies). The proposed anthology is discussed in the article "Drawing Knowledge from Information: Early Modern Texts and Images on the TAPoR Platform" [1].

The radical changes undergone by the institution of marriage in France during and after the Counter Reformation generated texts of several different genres. Included in the anthology will be medical, legal, religious, satirical and literary documents and engravings, all heavily annotated. It is the engravings that interest us for this presentation.

As part of a prototype for the anthology, several verse and prose polemics against marriage were encoded with XML markup in 2004 and early 2005. Most engravings of the period whose subject is marriage also fall into the polemical or satirical genre. Six from the collection of

the Cabinet des Estampes of the Bibliothèque Nationale de France were requested on CD Rom, and are the test images for this project:

Jacques Lagniet, "Il cherche ce qu'il voudrait ne pas trouver"
Abraham Bosse, "La femme qui bât son mari"
Jean Le Pautre, "Corrige, si tu peux, par un discours honneste"
François Guérard, "Le Grand Bureau ou la confrèrie des martires"
Nicolas Guérard, "Présage malheureux"
Nicolas Guérard, "Argent fait tout" [2]

THE SCHEMA

Marking up annotations in XML required a framework that allowed for both the text of the annotations, and the image areas to which they correspond, to be encoded in a single document. Given that well-established tagsets exist for each of these functions, an XML model was developed based on a marriage of the Scalable Vector Graphics (SVG) 1.1 specification [3], and a subset of the Text Encoding Initiative (TEI) P5 guidelines [4]. This union allows TEI and SVG markup to operate concurrently. The TEI markup forms the overarching structure of a document, while elements belonging to the SVG vocabulary may appear in specific locations within the TEI encoding.

Elements belonging to the SVG vocabulary are permitted within [div] tags and must be enclosed by the [svg] root element. Therefore, [svg] elements may appear anywhere with the TEI markup where [div] elements are permitted. Within an [svg] element may be any number of [rect] elements whose coordinates demarcate the area on an associated image to which an annotation applies. The texts of the annotations are enclosed in [div] blocks of their own, separate from the SVG encoding. This allows for annotation text to be encoded in any TEI-conformant way, presenting the possibility of integrating annotations with larger corpora. A [div] element containing an annotation text is associated explicitly with a set of annotation coordinates through references to the coordinates' svg:id attribute.

Markup validity is enforced through XML Schema or RELAX NG schema files which are bundled with the application. The schema which validates the TEI portion of the encoding has been generated by the ROMA suite of tools provided by the TEI for the purposes of specifying and documenting a customization. The TEI schema is supplemented by the addition of a schema describing the SVG 1.1 specification. The W3C provides the SVG 1.1 schema in either RELAX NG or DTD format, from which an XML Schema version may be derived using Trang [5]. Integrating schema from two different tagsets in this way is greatly facilitated by the modular construction inherent to both the TEI and SVG schema models. SVG may be 'plugged-in' to TEI by adding the [svg] root element to the list of allowable content in a particular context, and then associating the requisite schema documents with one another for the purposes of validation.

Taking this approach to schema marriage has several advantages. The TEI guidelines for textual encoding provide a tagset whose usage rules are well-defined and understood, facilitating the portability of the encoding between projects, and easing the integration of corpora from different sources. An earlier method of encoding image annotations in XML, Image Markup Language [6], is based on a standalone markup structure which does not offer the same high degree of interoperability as the current model. The TEI encourages customization of its guidelines to accommodate for a wide range of implementations, an approach this project demonstrates. More generally, working with XML allows for the encoded material to be transformed into other formats as requirements dictate, such as XHTML, PDF, or OpenDocument format.

THE IMAGE MARKUP TOOL

Having decided on our approach to a schema, we then began to look at how we might create the markup. We wanted a straightforward tool for defining areas in an image and associating them with annotative markup, and we looked initially at two possible existing tools, INote [7] and the Edition Production Technology [8].

INote, from the University of Virginia, is a Java application for annotating images. It does not appear to have been updated since 1998. In some ways, INote is an ideal tool; it is cross-platform (written in Java), and covers most of our requirements. However, we rejected INote for several reasons. The program can load only GIF and JPEG images, and we wanted to be able to handle other common image formats such as BMP and PNG. INote also allows only limited zooming (actual size, double size, and half size). We required more flexible

zooming to handle larger images. Finally, INote uses a proprietary file format.

However, INote does allow for polygonal and elliptical annotation areas, something not yet implemented in our own tool.

The Edition Production Technology (EPT) platform is an Eclipse-based software suite developed by the ARCHway project [9]. Its ImagText plugin allows the association of rectangular areas of an image with sections of transcribed text. Although it promises to be a very powerful tool, especially for the specific job of associating document scans with transcription text, the interface of the program is complex and would be confusing for novice users. In addition, the tool developers expect and encourage the use of customized DTDs ("We do not provide support or guarantees for the DTDs included in the demo release - it is expected that users will provide their own DTDs and thus their own specific encoding practices." [10]) The EPT also supports only JPEG, GIF, TIFF, and BMP files; other formats such as PNG are not supported ([http://rch01.rch.uky.edu/~ept/Tutorial/preparing_files.htm#images]).

We therefore decided to write our own markup program, which is called the Image Markup Tool [11].

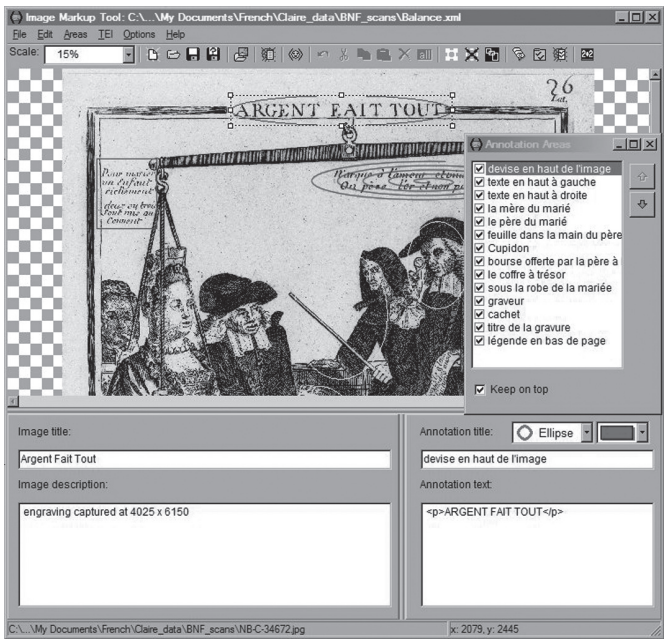


Fig 1: scrshot_main_1.jpg, avalable at [http://mustard.tapor.uvic.ca/~mholmes/image_markup/scrshot_main_1.jpg]

At the time of writing, the program is in at the «alpha» stage, and the first public version will be released under an open-source licence in December 2005. The program is written in Borland Delphi 2005 for Windows 2000 / XP. Development of the tool is guided by the following requirements:

The Image Markup Tool should:

- be simple for novices to use
- load and display a wide variety of different image formats
- allow the user to specify arbitrary rectangles on the image, and associate them with annotations
- allow such rectangles to overlap if the user wishes
- provide mechanisms for bringing overlapped rectangles to the front easily
- require no significant knowledge of XML or TEI
- allow the insertion of XML code if the user wishes
- save data in an XML file which conforms to a TEI P5-based schema with embedded SVG
- reload data from its own files
- come packaged with an installer, Help file, and basic tutorial

Using the Image Markup Tool, we have been able to perform several types of direct annotations, including the text within the engravings, commentary on that text, commentary on significant gestures depicted, and information about the engraver and the seal of the library at the time the engraving entered the library's collection. The tool allows for distinction among types of annotation, and the use of a TEI-based file format allows us to link easily between the markup of the engravings the TEI-encoded polemical texts which are also included in the collection.

We are now planning to use the program for a future project which involves marking up scans of historical architectural plans. One of the aims of this project will be to make the plans available to the public, so that (for example) the current owners of heritage buildings will be able to do renovation and restoration work with more detailed knowledge of the original building plan.

References

[1] **Carlin, Claire.** *"Drawing Knowledge from Information: Eary Modern Texts and Images on the TAPoR Platform"* in *Working Papers from the First and Second Canadian Symposium on Text Analysis Research* (CaSTA), [http://www2.arts.ubc.ca/chwp/Casta02/], and forthcoming in *Text Technology*.

[2] Bibliothèque nationale (France). Département des estampes et de la photographie and Roger-Armand Weigert, ed. *Inventaire du fonds français, graveurs du XVIIe siècle*. Paris: Bibliothèque Nationale, 1939.

[3] **Ferraiolo, Jon, et al., eds.** (2003). *Scalable Vector Graphics* (SVG) 1.1 Specification. World Wide Web Consortium. [http://www.w3.org/TR/SVG/] [Accessed 03-11-2005].

[4] **Sperberg-McQueen, C. M., and Lou Burnard, eds.** (2005). *The TEI Guidelines: Guidelines for Electronic Text Encoding and Interchange - P5*. The TEI Consortium. [http://www.tei-c.org/P5/Guidelines/index.html] [Accessed 03-11-2005].

[5] **Thai Open Source Software Center.** (2003). *Trang - Multi-format schema converter based on RELAX NG*. [http://www.thaiopensource.com/relaxng/trang.html] [Accessed 03-11-2005].

[6] **Image Markup Language 1.0.** (2000). [http://faculty.washington.edu/lober/iml/] [Accessed 03-11-2005].

[7] _INote_. Intitute for Advanced Technology in the Humanities, University of Virginia, 1998. [http://www.iath.virginia.edu/inote/] [Accessed 07-11-2005]

[8] *_Edition Production Technology* (EPT build 20050301). **Kiernan, Kevin et al.**, University of Kentucky. 2005. [http://rch01.rch.uky.edu/~ept/download/] [Accessed 07-11-2005]

[9] **Kiernan, Kevin, et al.** *The ARCHway Project*. University of Kentucky. [http://beowulf.engl.uky.edu/~kiernan/ARCHway/entrance.htm] [Accessed 07-11-2005]

[10] «Tutorial for EPT Demonstration.» [http://rch01.rch.uky.edu/~ept/Tutorial/demo_tagging.htm] [Accessed 07-11-2005]

[11] *Image Markup Tool*. **Carlin, Claire, Eric Haswell and Martin Holmes.** University of Victoria Humanities Computing and Media Centre. 2005. [http://mustard.tapor.uvic.ca/~mholmes/image_markup/]

Methods for Genre Analysis Applied to Formal Scientific Writing

Paul CHASE
chaspau@iit.edu

Shlomo ARGAMON
argamon@iit.edu

Linguistic Cognition Laboratory Dept. of Computer Science Illinois Institute of Technology 10 W 31st Street Chicago, IL 60616, USA

1 Overview

Genre and its relation to textual style has long been studied, but only recently has it been a candidate for computational analysis. In this paper, we apply computational stylistics techniques to the study of genre, which allows us to analyze large amounts of text efficiently. Such techniques enable us to compare rhetorical styles between different genres; in particular, we are studying the communication of scientists through their publications in peer-reviewed journals. Our work examines possible genre/stylistic distinctions between articles in different fields of science, and seeks to relate them to methodological differences between the fields.

We follow Cleland's (2002) work in this area and divide the sciences broadly into Experimental and Historical sciences. According to this and other work in the philosophy of science, Experimental science attempts to formulate general predictive laws, and so relies on repeatable series of controlled experiments that test specific hypotheses (Diamond 2002), whereas Historical science deals more with contingent phenomena (Mayr 1976), studying unique events in the past in an attempt to find unifying explanations for their effects. We consider the four fundamental dimensions outlined by Diamond (2002, pp. 420-424):

1. Is the goal of the research to find general laws or statements or ultimate (and contingent) causes?
2. Is evidence gathered by manipulation or by observation?
3. Is research quality measured by accurate prediction or effective explanation?
4. Are the objects of study uniform entities (which are interchangeable) or are they complex entities (which are ultimately unique)?

The present experiment was designed to see if language features support these philosophical points. These linguistic features should be topic independent and representative of the underlying methodology; we are seeking textual clues to the actual techniques used by the writers of these scientific papers. This paper is partially based on our previously presented results (Argamon, Chase & Dodick, 2005).

2 Methodology

2.1 The Corpus

Our corpus for this study is a collection of recent (2003) articles drawn from twelve peer-reviewed journals in six fields, as given in Table 1. The journals were selected based both on their prominence in their respective fields as well as our ability to access them electronically, with two journals chosen per field and three fields chosen from each of Historical and Experimental sciences. Each article was prepared by automatically removing images, equations, titles, headings, captions, and references, converting each into a simple text file for further processing.

2.2 Systemic Functional Linguistics

We base our analysis on the theory of Systemic Functional Linguistics (SFL; Halliday 1994), which construes language as a set of interlocking choices or systems for expressing meanings, with general choices constraining the possible more specific choices. SFL presents a large number of systems, each representing a certain type of functional meaning for a potential utterance. Each system has conditions constraining its use and several options; once within a system we can choose but one option. Specific utterances are constrained by all the systemic options they realize. This approach to language allows the following types of questions to be asked: In places where a meaning of general type A is to be expressed in a text, what sorts of more specific meanings are more likely to be expressed in different contexts?

We focused on several systems for this study, chosen to correspond with the posited differences between the types of science we study: Expansion, Modality, and Comment (Matthiessen 1995). Expansion describes features linking clauses causally or logically, tying in to dimensions 1 and 4 above. Its three types are: Extension, linking different pieces of information; Elaboration, deepening a given meaning via clarification or exemplification; and Enhancement, qualifying previous information by spatial, temporal, or other circumstance. The second system, Modality, relates to how the likelihood, typicality, or necessity of an event is indicated, usually by a modal auxiliary verb or an adjunct adverbial group; as such it may serve to indicated differences on dimensions 2, 3, and 4. There are two main types of modality: Modalization, which quantifies levels of likelihood or frequency, and Modulation, which qualifies ability, possibility, obligation, or necessity of an action or event. Finally, the system of Comment is one of assessment, comprising a variety of types of ``comment'' on a message, assessing the writer's attitude towards it, its validity or its evidential status; this provides particular information related to dimensions 1 and 3.

In our analysis, it will be most helpful to look at oppositions, in which an option in a particular system is strongly indicative of one article class (either Experimental or Historical science) while a different option of that same system is indicative of the other class. Such an opposition indicates a meaningful linguistic difference between the classes of articles, in that each prefers a distinctive way (its preferred option) of expressing the same general meaning.

2.3 Computational analysis

Because hand analysis is impractical on large document sets the first analyses were done via computer. We built a collection of keywords and phrases indicating each option in the aforementioned systems. Each document is first represented by a numerical vector corresponding to the relative frequencies of each option within each system. From here, machine learning was applied in the form of the SMO (Platt 1998) algorithm as implemented on the Weka machine learning toolkit (Witten & Frank 1999), using 10-fold cross-validation in order to evaluate classification effectiveness. This method was chosen in part because it generates weights for each feature; a feature has high weight (either positive or negative) if it is strongly indicative for one or the other class.

2.4 Human annotation

To measure the validity of our computational analysis, we are also performing hand tagging of systemic features on a subset of the corpus articles. Two articles from each journal have been chosen, each to be tagged by two trained raters. Part of the tagging process is to highlight key words or phrases indicating each option; we will compare these statistics to our previously generated feature lists in order to test and refine them. The tagging is currently under way; we will present results at the conference.

4 Results

To determine the distinctiveness of Historical and Experimental scientific writing, the machine learning techniques described above were applied to pairs of journals, giving for each pair a classification accuracy indicating how distinguishable one journal was from the other. These results are shown in Figure 1, divided into four subsets: Same, where both journals are from the same science; Hist and Exper with pairs of journals from different sciences, but the same type; and Diff indicates pairings of Historical journals with Experimental ones. The thick black line indicates the mean for each set, and the outlined box represents the standard deviation. As we see, journal pairs become more distinguishable as their methodological differences increase. Interestingly, Historical journals appear more stylistically homogenous than the Experimental journals, which is a subject for further study.

This shows that SFL is capable of discriminating between the different genres presented. We also examined the most important features across the 36 trials between different journals. The most consistently indicative-those features that are ranked highest for a class in at least 25 trials-are presented in Table 2. The table is arranged as a series of oppositions: the features on each row are in the same system, one side indicating Historical, the other Experimental.

In the system of Expansion, we see an opposition of Extension and Enhancement for Historical and Experimental sciences, respectively. This implies more independent information units in Historical science, and more focused storylines within Experimental science. Furthermore, there are oppositions inside both systems, indicating a preference for contrasting information (Adversative) and contextualization (Matter) in Historical science and for supplementary Information (Additive) and time-space (Spatiotemporal) relations in Experimental science.

The system of Comment also supports the posited differences in the sciences. The Experimental sciences' preference for Predictive comments follows directly from their focus on predictive accuracy. On the Historical side, Admissive comments indicate opinions (as opposed to factual claims), similarly Validative comments show a concern with qualifying the validity of assertions, comprising more of strong evidence than rigid proofs.

Finally in Modality we see interesting contrasted features. On the top level we have near-perfect opposition between Modalization and Modulation in general; Historical sciences speak of what is 'normal' or 'likely', while Experimental sciences assess what 'must' or 'is able' to happen.

5 Conclusion

This work is the first step in developing new automated tools for genre analysis, which promises the possibility of automatically analyzing large corpora efficiently or stylistic aspects while giving human interpretable results. The specific research presented has implications for the understanding of the relationship between scientific methodology and its linguistic realizations, and may also have some impact on science education. Future work (beyond the hand annotation and analysis already in progress) includes looking into stylistic variation within different article sections, as well as other analysis techniques (such as principle components analysis).

Journal	#Art	Avg. Words
J. Geology	93	4891
J. Metamorphic Geol.	108	5024
Biol. J. Linnean Society	191	4895
Human Evolution	169	4223
Palaeontologia Electronica	111	4132
Quaternary Research	113	2939
Physics Letters A	132	2339
Physical Review Letters	114	2545
J. Physical Chemistry A	121	4865
J. Physical Chemistry B	71	5269
Heterocycles	231	3580
Tetrahedron	151	5057

Table 1: Journals used in the study; the top represents historical fields with experimental sciences below.

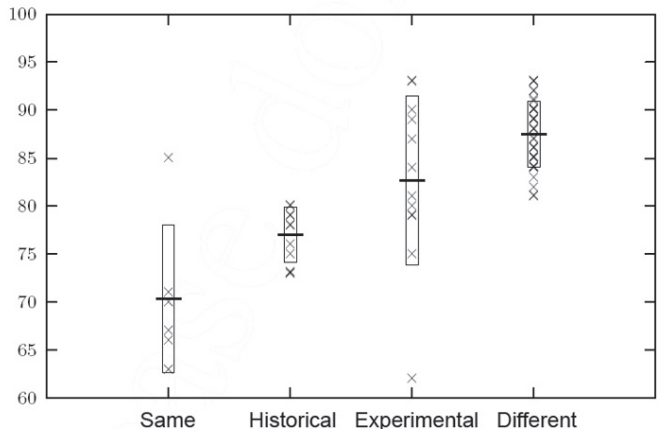


Figure 1: Learning accuracy for distinguishing articles in different pairs of journals. 'Same' are pairs where both journals are in the same field, 'Historical' and 'Experimental' represent pairs of journals in different Historical and Experimental fields, and 'Different' pairs of journals where one journal is experimental and the other historical. Means and standard deviation ranges are shown.

System	Historical	Experimental
Expansion	Extension(26)	Enhancement(31)
Elaboration		Apposition(28)
Extension	Adversative(30)	Additive(26)
Enhancement	Matter(29)	Spatiotemporal(26)
Comment	Admissive(30) Validative(32)	Predictive(36)
Modality Type	Modalization(36)	Modulation(35)
Modulation	Obligation(29)	Readiness(26)
Modality Value		High(27)
Modility Orientation	Objective(31)	Subjective(31)

Table 2. Consistent indicator features within each of the systems used in the study. Numbers in parentheses show in how many paired-classification tests the feature names was an indicator for the given class of documents.

References

Argamon, S., Chase, P., and Dodick, J.T. (2005). The *Languages of Science: A Corpus-Based Study of*

Experimental and Historical Science Articles. In Proc. 27th Annual Cognitive Science Society Meetings.

Baker, V.R. (1996). *The pragmatic routes of American Quaternary geology and geomorphology*. Geomorphology 16, pp. 197-215.

Cleland, C.E. (2002). *Methodological and epistemic differences between historical science and experimental science*. Philosophy of Science.

Diamond, J. (2002). *Guns, Germs, & Steel*. (New York: W. W. Norton and Company).

Halliday, M.A.K. (1991). *Corpus linguistics and probabilistic grammar*. In Karin Aijmer & Bengt Altenberg (ed.), English Corpus Linguistics: Studies in honour of Jan Svartvik. (London: Longman), pp. 30-44.

Halliday, M.A.K. (1994). *An Introduction to Functional Grammar*. (London: Edward Arnold).

Halliday, M. A. K., & R. Hasan. (1976). *Cohesion in english*. London: Longman.

Halliday, M. A. K., & J.R. Martin. (1993). *Writing science:Literacy and discursive power*. London: Falmer

Joachims, T. (1998). *Text categorization with Support Vector Machines: Learning with many relevant features*. In ECML-98,10th European Conference on Machine Learning, pp. 137-142.

Mayr, E. (1976). *Evolution and the Diversity of Life*. (Cambridge: Harvard University Press).

Mitchell, T. (1997) *Machine Learning*. (McGraw Hill).

Platt, J. (1998), *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,* Microsoft Research Technical Report MSR-TR-98-14.

Witten, I.H. and Frank E. (1999). *Weka 3: Machine Learning Software in Java*: http://www.cs.waikato.ac.nz/~ml/weka.

Combining Cognitive Stylistics and Computational Stylistics

Louisa CONNORS

School of Humanities and Social Science
The University of Newcastle, Australia

Studies in the computational analysis of texts have been successful in distinguishing between authors and in linking anonymously published texts with their authors but computational tools have yet to be accepted as mainstream techniques in literary analysis. Criticisms are generally centred around the belief that computational analyses make false claims about scientific objectivity and are in fact no less subjective than any other critical approach. This is perhaps because computational projects are in conflict, at a fundamental level, with contemporary post-structuralist notions of subjectivity, meaning and the arbitrary nature of language. This paper will argue that these objections rest on assumptions about language that need to be examined in light of developments in linguistics and cognitive psychology and that cognitive linguistics has the potential to bring a more interpretive framework to computational stylistics, a practice that has traditionally been applied in fairly narrow, empirical way.

Whilst computational analysis points to the possibility of subjectivity that is more coherent than some theoretical approaches imply, it does not necessarily diminish the role of culture and context in the formation of texts and subjectivity as highlighted by materialist readings. The application of cognitive linguistics in a computational study provides a model of syntax and semantics which is not independent of context but deeply bound up in context. Cognitive linguistics can explain the existence of computational results in a way that Saussurean based theories can not. It can offer a rich interpretive model that does not neglect the importance of author, reader, or context through its approach to language and literature as an expression of an innately constrained and embodied human mind.

Computational stylistics, particularly in studies of attribution, generally makes use of function words in

order to distinguish between texts. As Craig (2004) explains, independent variables like genre, are compared with counts of internal features, or dependent variables, like function words. Correlation of these two kinds of variables is the primary tool of computational stylistics (275-76). Critics of stylistics, most notably Stanley Fish, tend to privilege individual instances of particular words in interpretive communities over more general rules, and question the validity of the stylistic project. From a cognitive perspective, however, language possesses universal features because it emerges from the interaction of "inherent and experiential factors" that are "physical, biological, behavioural, psychological, social, cultural and communicative" (Langacker 1). Langacker claims that each language "represents a unique adaptation to common constraints and pressures as well as to the peculiarities of its own circumstances" (1). Computational stylistics of the kind undertaken in this study provides us with evidence of the peculiarities and creative adaptations of an individual user, and also highlights more general trends which can be used for comparative purposes.

Our attitude to what we can say about a text depends largely on our account of language. Widely shared post-structuralist assumptions about language and indeterminacy have contributed to the lukewarm reception of computational stylistics in literary interpretation. In cognitive linguistics "Semantics is constrained by our models of ourselves and our worlds. We have models of up and down that are based on the way our bodies actually function. Once the word "up" is given its meaning relative to our experience with gravity, it is not free to "slip" into its opposite. "Up" means up and not down" (Turner 7). Cognitive stylistics views a text as the product of a human action and it therefore carries the mark of that action. The cognitive belief that language and conventional thought emerge from "our perception of a self within a body as it interacts with an environment" suggests that meaning is somewhat constrained and that "some form of agency is fundamental to language" (Crane 22).

The idea of authorial agency is one that is rejected by structuralist and post-structuralist critics. In proclaiming the death of the author Barthes suggests that the text becomes an 'open sea', a space of 'manifestly relative significations, no longer tricked out in the colors of an eternal nature' (Barthes 170). The notion of the "transcendental signified" rejected by post-structuralist critiques is not, however, the notion of agency proposed by cognitive stylistics. The view of the "Author" rejected by Barthes, Derrida and Foucault is as, Seán Burke explains "a metaphysical abstraction, a Platonic type, a fiction of the absolute" (27). Cognitive stylistics tends not to deal in absolutes.

Stylistics is one way of getting evidence and making sense of texts as human actions. Through its approach to thought and language, cognitive stylistics points to issues that are of concern to scholars of literature, such as "subject formation, language acquisition, agency and rhetoricity" (Richardson 157). Cognitive philosophy claims that the mind is embodied and that concepts are therefore created "as a result of the way the brain and body are structured and the way they function in interpersonal relations and in the physical world" (Lakoff and Johnson 37). The links between the brain and the body mean an objective reality is impossible given the role our sensorimotor system plays in perception. But as Lakoff and Johnson explain, it is our sensorimotor system's role in shaping conceptual systems that keeps these systems in touch with the world (44).

The embodied cognitivism of Lakoff and Johnson, also known as second generation cognitivism, argues that our access to the external world is mediated through cognitive processes. Cognitivism provides a framework in which we can still legitimately engage with psychoanalytic interpretations, gender focused readings, and the material conditions of production while using computational and cognitive techniques of analysis. Computers enable us to draw together instances of common forms, and other features of a text, in a way that would be simply impossible to an individual human reader. Cognitive linguistics provides a theoretical justification for paying attention to common forms in the first place, and reveals a way in which the features highlighted by computational approaches can contribute something of value to traditional literary analysis.

References

Barthes, Roland. *On Racine*. Trans. Richard Howard. New York: Octagon Books, 1977.

Burke, Seán. *The Death and Return of the Author:*

Criticism and Subjectivity in Barthes, Foucault and Derrida. Edinburgh: Edinburgh University Press, 1988.

Craig, Hugh. "Stylistic Analysis and Authorship Studies." *A Companion to Digital Humanities*. Eds. Susan Schreibman, Ray Siemens and John Unsworth: Blackwell, 2004. 273-88.

Crane, Mary Thomas. *Shakespeare's Brain: Reading with Cognitive Theory*. Princeton and Oxford: Princeton University Press, 2001.

Lakoff, George, and Mark Johnson. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books, 1999.

Langacker, Ronald W. *Foundations of Cognitive Grammar: Descriptive Applications*. Vol. II. Stanford, California: Stanford University Press, 1991.

Richardson, Alan. "Cognitive Science and the Future of Literary Studies." Philosophy and Literature 23.1 (1999): 157-73.Turner, Mark. Death Is the Mother of Beauty. Chicago: University of Chicago Press, 1987.

The Introduction of Word Types and Lemmas in Novels, Short Stories and Their Translations

Mária CSERNOCH

University of Debrecen, Hungary

Introduction

In earlier analyses of the introduction of word types in literary works authors came to contradictory conclusions. Some suggested, in accordance with reader's intuition, that the launch of new chapters and a sudden increase in the number of the newly introduced word types (NWT) usually coincide. Others, on the other hand, found that there is no clear connection in the rise of NWT and the beginning of chapters, rather an increase in NWT appears at longish descriptions with rather stylistic reasons.

Words do not occur randomly in texts, so the ultimate goal of building models based on word frequency distributions may not be the reproduction of the original text. Nevertheless, models based on the randomness assumption give reliable information about the structure of the texts (for review see Oakes, 1998; Baayen, 2001). To further our knowledge in this field the source of the bias between the original and the model-based texts should be examined. Baayen (1996; 2001) described a systematic overestimation for the expected vocabulary size and found that this bias disappears when the order of the sentences is randomized, indicating that the bias should not be attributed to constrains operating on sentence level.

To prove that this misfit is due to significant changes on discourse level we introduced several new concepts during the process of building the model and analyzing the results (Csernoch, 2003). Among these the fundamental step was to scrutinize NWT in hundred-token-long intervals rather than examining the overall vocabulary size. Next, instead of eliminating the bias between these artificial texts and original works, the significant protuberances on the graphs of NWT were

examined. First monolingual sets of works were processed then, to improve the comparison we also analyzed original English texts and their Hungarian translations together with English and German translations of a Hungarian text.

Assuming that the changes occur on discourse level, the language in which the text is written should have no significance. In other words, neither syntactic nor semantic constrains on sentence or paragraph level should matter, and only events that occur on discourse level will provide substantial alterations in the flow of the text, and thus produce considerable protuberances on the graphs of NWT.

Methods

Building the model

To analyze a text first the number of different word types was counted, the frequency of each was determined, and then based on these frequencies a dynamic model was built (Csernoch, 2003). The model generated an artificial text whose word types had the same frequencies as in the original text and was able to reproduce the trends of the original text. However, changes which are only seasonal – protuberances – did not appear in the artificial text. To locate these protuberances the difference between the original and the model text was calculated. We then determined the mean (M) and the standard deviation (SD) of the difference. Protuberances exceeding M±2SD were considered significant.

The distribution of the hapax legomena was also examined. Assuming that they are binomially distributed their expected mean (M_h) and standard deviation (SD_h) were calculated and again those points where considered significant which exceeded M_h+2SD_h.

Original texts compared to their translations

Original texts were not only compared to the model-generated artificial texts but to their translations in other natural languages. In this study we analyzed the Hungarian novel, SORSTALANSÁG from Imre Kertész and its English (FATELESS) and German (ROMAN EINES SCHICKSALLOSEN) translations, Rudyard Kipling's THE JUNGLE BOOKS and their Hungarian translations (A DZSUNGEL KÖNYVE), and Lewis Caroll's ALICE ADVENTURES IN WONDERLAND and THROUGH THE LOOKING GLASS and their Hungarian translations (ALICE CSODAORSZÁGBAN and ALICE TÜKÖRORSZÁGBAN).

These three languages were chosen because they are different in their morphological structures, it is hard to trace any common syntactic characteristic which all three share.

Analyzing lemmatized texts
To check whether the analyses of the raw, un-lemmatized texts give reliable information for the introduction of NWT the lemmatization of both the English and the Hungarian texts was carried out. The English texts were tagged and lemmatized by CLAWS (the Constituent Likelihood Automatic Word-tagging System) [1], while the morphological analysis of the Hungarian texts was carried out by Humor and the disambiguation was based on a TnT tagger [2].

Results

Comparing the texts and their translations it was first found that the morphologically productive Hungarian texts had the smallest number of running words and lemmas while the largest number of hapax legomena both in the lemmatized and un-lemmatized versions. In contrast, the English texts contained the most running words but the smallest number of hapax legomena.

To each text and language an individual model was created. Based on these models the positions of the significant protuberances were traced and compared to each other in the original texts and their translations. It was noticed that regardless of the actual language these protuberances occurred in most cases at the same position, that is, at the same event in the flow of the story.

We could clearly establish that the protuberances were found at places where new, usually only marginally connected pieces of information were inserted into the text rather than at new chapters. This idea was strengthened by a peculiarity of the English translation of SORSTALANSÁG, namely that the boundaries of chapters are different from those of the Hungarian and German texts, which further substantiates that the protuberances do neither necessarily coincide with the beginning nor are hallmarks of a new chapter. Similarly, in the original Alice stories the boundaries of the chapters are eliminated by unusual

typographic tools, while in the Hungarian translation these boundaries are set back to normal. Neither the English nor the Hungarian texts produced any protuberances at these places. In THE JUNGLE BOOKS we again found that the significant differences between the original text and the model are not necessarily at the beginning of a new tale, except for cases when a new setting is introduced.

The fact that these descriptions have only a stylistic role in the text was further substantiated by examining the distribution of hapax legomena. The number of hapax legomena was found to be high exactly at the same positions of the text where protuberances in the number of the newly introduced word types occurred.

To examine the lemmatized version of the texts carried some risk since loosing the affixes might eliminate the change in mode, time, style, etc., while, on the other hand, might reveal events lost in word types carrying the affixes. Since our dynamic model is capable of giving a relatively good estimation for the introduction of words, the question was whether using lemmas instead of word types would provide additional information gained by comparing the artificial texts and the translations to the original text.

In the English texts the lemmatization did not reveal any additional information, the protuberances occurred at exactly the same places in the lemmatized as in the un-lemmatized versions. In un-lemmatized Hungarian texts the first protuberance usually occurred later than in corresponding English and German texts, although we were able to locate them by examining protuberances that were somewhat below the level of significance. In these cases lemmatization helped, and we got clear protuberances reaching the level of significance in lemmatized Hungarian texts.

The comparison of the dynamic model built to lemmatized texts in different languages might also be used to analyze and compare the vocabulary of the original texts and their translations. It would, furthermore, enable the comparison of the stylistic tools used by the original author and the translator in the introduction of new words.

Summary

Using lexical statistical models for analyzing texts the explanation for the difference between the original and the model-based artificial text was examined. It was found that changes on discourse rather than on sentence or paragraph levels are responsible for these differences. Two methods were used to prove this. First, texts and their translations, both lemmatized and un-lemmatized versions, were analyzed and compared to a dynamic model built on the randomness assumption to find that the significant changes on the graphs of the newly introduced word types occurred at corresponding positions within the translations. Second, the distribution of hapax legomena was compared to a binomial distribution, again to find that the significant differences between the original and the predicted distributions occurred at descriptions, only in loose connection with the antecedents and what follows. More importantly, these coincided with the significant changes of the newly introduced word types.

References

Baayen, R. H. (1996) *The Effect of Lexical Specialization on the Growth Curve of the Vocabulary.* Computational Linguistics 22. 455-480.

Baayen, R. H. (2001) *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht, Netherlands

Csernoch, M. (2003) *Another Method to Analyze the Introduction of Word-Types in Literary Works and Textbooks*. Conference Abstract, The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities Göteborg University, Sweden

Oakes, M. P. (1998) *Statistics for Corpus Linguistics*. Edinburgh University Press

[1] http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/

[2] http://corpus.nytud.hu/mnsz/index_eng.html

Digitizing Cultural Discourses. Baedeker Travel Guides 1875 - 1914.

Ulrike CZEITSCHNER

Ulrike Czeitschner, AAC-Austrian Academy Corpus

The Baedeker project, presented in this proposal, is initiated within the framework of the Austrian Academy Corpus, a research department at the Austrian Academy of Sciences. The sub corpus comprises first editions of German travel guidebooks brought out by the Baedeker publishing house between 1845 and 1914. The texts cover exclusively non-European destinations. Today they are rare, which is due to low print runs.

The aim of the project is threefold: 1) it partly deals with the genre from a literary point of view, 2) it looks at travel guides as cultural historical resources as well as artefacts that represent various discourses on culture, and 3) it examines the capacities of digital resources.

Travel guidebooks, not regarded as a literary form in their own right within the classical canon, played a minor role in comparison with travel narratives for a long time. Substantial contributions to the historical development of the genre and its specific language of expression are still few. Even postcolonial literary criticism cannot be regarded as an exception in this respect. Surprisingly enough, few approaches appreciate the importance and impact of this genre on the establishment and maintenance of Orientalist discourses and colonial practises. Linguistic accounts mostly concentrate on contemporary travel guides and exclude historical development and change. Thus, digital versions of early travel guidebooks can provide an incentive to improve both comparative linguistic and literary genre studies.

Furthermore, research on tourism history, its influence on modern society, its bearing on social and cultural change has increased quantitatively in many fields of the humanities since the 1980s: history in general, art history, colonial studies, social and cultural anthropology, economics, geography and tourism studies, today an accepted sub branch of sociology. In these disciplines it goes without saying that travel guidebooks are valuable sources. Nonetheless, they are often dealt with as sources among many others. (The few exceptions are e.g. James Buzard 1993; Kathleen R. Epelde 2004, Sabine Gorsemann 1995, Rudy Koshar 2000 and the research group Türschau 16 1998.)

The Baedeker, appearing from 1832 onwards, set the standard and defeated all competition both inside and outside the German-speaking countries. One cannot talk about travel guides without Baedeker coming to mind. During the first decades of the Baedekers, the focus was on Europe. However, they were issued for non-European destinations as well, an aspect missing from critical literature. The guides in the Baedeker-Corpus cover a variety of regions such as Palestine and Syria (1875), Lower and Upper Egypt (1877, 1891), North America and Mexico (1883), Asia Minor (1905), the Mediterranean coastline of Africa (1909) as well as India (1914). Dealing with a wide range of cultural environments the "Tourist Gaze" upon the "Other" has to be scrutinized in greater depth. Assuming that images of the "Other" reflect cultural self-perception to a great extent, travel guidebooks tell at least as much about the "Self" as about the "Other". For well known reasons, none of the components involved here can be taken for granted as precise, unambiguous or fixed and independent entities. Taking this argument seriously, self-images - like all the other components - are to be understood as flexible phenomena. Moving away from the very frequent restriction on one region or country the Baedeker project turns towards a wider geographical diversification to explore the German repertoire of how one used to speak at the turn of the 19th century about one's own culture, and at the same time, that of others.

As concerns the digital methods by which these phenomena are to be investigated, the following has to be pointed out: while XML is now an accepted standard for the creation and exchange of digital data, it is necessary to move towards a closer consideration of domain-specific XML vocabularies. All Baedekers have undergone scanning, OCR and basic XML annotation as usual with all AAC projects. The task at hand is to devise a schema to markup the features of the Baedekers, relevant for their role in travel history. Existing standards like AnthML and Archaeological Markup Language are focused on material artefacts. A markup language

which considers immaterial cultural aspects is missing so far. As a matter of course the development of a standardized language needs the expertise of the wider scholarly community and has to be a team effort - a requirement not feasible within one institution. Thus, the Baedeker project should be seen as a small contribution in preparing such a markup language, designing a sub-set of tags focusing on a well defined segment of cultural life.

The main challenge is encoding what travel guides are essentially supposed to do, namely introducing foreign cultures and people/s, recommending an itinerary, assessing sites and festivals, cultural and social conditions, suggesting modes and attitudes of behaviour to adopt in these places and on these occasions. Since cultural knowledge as well as recommendations, valuations, stereotypes or comparisons often is articulated in an implicit manner, which is difficult to encode, the project targets subject-matters such as people/s, languages and religions, social, political, and other cultural concepts as well as sights being recommended, valuated, stereotyped, and compared in the travel guides. As an example I will refer to the repertoire of "group designations", be it ethnic, national, social, religious, political, and occupational, showing how they relate to the historical discourse on culture. The paper will demonstrate that subject-matters can be easily marked up, they allow for an appropriate access to context - i.e. different routes to topics and explicit as well as implicit knowledge - and they provide a basis for comparative analysis. In addition, this strategy separates annotation from interpretation and limits the risk of encoding preconceived assumptions.

Detailed domain-specific markup can be applied to other texts dealing with similar topics - to primary and secondary sources, historical as well as contemporary. In this respect the Baedeker can be seen as a starting point. Using open standards allows for ongoing quality enhancement and adjustment. XML annotation, in this sense, is not a single-serving tool, but a permanent enrichment, accessible and shareable with the wider scholarly community. Retrieval results can be reviewed by different scholars, paving the way for reinterpretation and new questions. It is expected that differing results will come from the same markup.

TELOTA - the Electronic Life of the Academy. New Approaches for the Digitization of Long Term Projects in the Humanities.

Alexander CZMIEL

Berlin-Brandenburg Academy of Sciences and Humanities

The Berlin-Brandenburg Academy of Science and Humanities (BBAW) [1] is an international and interdisciplinary association of excellent scholars with a distinctive humanities profile. The Academy hosts about 30 long term projects in the humanities with a project runtime often longer than 100 years. Examples of these projects include work on academic dictionaries, printed and manuscript editions, documentations, bibliographies, archival and publishing projects and more. These project groups have access to information and data made by outstanding scientists like Gottfried Wilhelm Leibniz, Johann Wolfgang von Goethe, Immanuel Kant, Albert Einstein, Jacob and Wilhelm Grimm, Alexander von Humboldt and many more who were members of the Academy during the last 300 years. Throughout its history the Society could rank 76 Nobel Laureates among its members.

According to the "Berlin Declaration on Open Access in the Sciences and Humanities" [2], which was signed by the President of the Academy, an initiative was founded to provide a sustainable, interactive and transparent way to inspire activities in supporting research, communication and presentation with electronic media or in other words to "electrify" the long term humanities projects hosted by the Academy. This initiative is called "TELOTA – the electronic life of the academy". [3]

One part of TELOTA is the so called "Project of the Month" (POM) [4] working group which started work in January 2005. The main task of this working group is to provide solutions for the mentioned issues which are cost efficient, future proven (especially to be independent

from commercial vendors) and freely extensible. Every month the data of a selected academy project is processed, (re-)structured and presented on the web to give researchers in the humanities, scholars and the interested public a new view in the extensive knowledge inventory of the academy. To identify and process information from the long term projects is one of the central tasks for POM. Further goals are:

1. To replace older, cost-intensive proprietary tools which are often not very suitable for presentations in the World Wide Web:
   - Improve the unification of the developed solutions for the different projects with respect for already existing solutions and open standards as well as concentration on third party open source software.
   - Production of reusable software modules.
2. To offer to the interested public an overview and provide an insight into the work done by the projects hosted by the academy:
   - To make new information and resources available on the World Wide Web.
   - Adaptation, unification and customization of existing data to offer a new point of view on a certain project.
   - Give access to the raw data which can be queried by arbitrary applications using XQuery. [5]
3. To benefit each humanities project hosted by the academy:
   - The projects should be able to access and administrate the data they produce on their own for a gradual extension of their web presence and research material.
   - Guarantee of long term accessibility and preservation as a result of consistent data and coherent administration.
   - Real time accessibility to the projects' results in the World Wide Web.
   - Support of the project's work flow with tools especially developed for their needs.

All the applied technologies and third party tools are reused, like all the gained experience is transferred from one project to the next. In addition new technologies are adopted to the working group's portfolio so it is able to react properly to the monthly changing requirements.

This paper will introduce the work of the "Project of the Month" working group and exemplary present two systems for humanities projects from the viewpoint of an "in-house" working group. It shows the possibilities of developing electronic resources of long term projects in a very short time period. Additionally it demonstrates a way how the mentioned technologies can be combined as flexible as possible.

The first system, the "scalable architecture for electronic editions", uses the opportunities of web services applied on critical text editions and was developed while processing prominent projects such as the "Corpus Medicorum Graecorum/Latinorum" or the "Marx-Engels Gesamtausgabe". The main component is a native XML-Database [6] which is able to interpret XQuery-scripts to form the web application. The user, according to his needs, dynamically decides on the view of the presented texts and translations and the information which is displayed like line numbers or links to the apparatus. So he can customize the electronic edition depending on his scientific position or interests. If possible, facsimiles are linked to text, translation and apparatus and if needed it is possible to search the electronic edition in different scripts, like ancient Greek.

The second system, an approach for digital dictionaries, shows the development stages of an interactive on-line dictionary which currently is work in progress and could contain the digital versions of dictionary projects of the academy in the future. Such a system is necessary for the real time digital presentation of dictionary project results. Examples are the "Dictionary of contemporary German language", the "Dictionary of Goethe" or the "German Dictionary of Jacob Grimm and Wilhelm Grimm". One main feature besides arbitrary querying the database using XQuery is the possibility to add and edit own dictionary articles, if the user is authorized to do so. The search results than can be displayed in HTML or PDF.

Both systems are conceptually designed with a general attempt but currently serve as sample applications. The use of a more technically matured version of this systems should not be limited to one project or just the academy

long term projects rather than potentially be open to any kind of critical text edition or dictionary.

References

[1] http://www.bbaw.de

[2] http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html

[3] http://www.bbaw.de/initiativen/telota/index.html

[4] http://pom.bbaw.de/index-en.html

[5] http://www.w3.org/TR/2005/CR-xquery-20051103/

[6] http://www.exist-db.org/

Eccentricity and / or Typicality in Sterne's Sermons ? Towards a Computer - Assisted Tentative Answer

Françoise DECONINCK-BROSSARD

Université Paris X, U.F.R. d'Etudes anglo-américaines

Laurence Sterne (1713-1768) is better-known nowadays as a novelist, but he was also an Anglican minister whose literary style was initially developed in the pulpit. He wrote many sermons before he turned to novel-writing. Indeed, his first publications were single sermons.

Melvyn New, who cannot be praised enough for publishing the first scholarly edition of his sermons, claims that it is "foolish" to "argue Sterne's uniqueness as a sermon-writer". Others have followed suit (e.g. Elizabeth Kraft). However, Paul Goring and Judith Hawley have cogently argued that Sterne's contemporaries often commented on the distinctiveness of Mr Yorick's sermons.

As a scholar who has devoted much of her research to eighteenth-century English pulpit literature, I have long had the impression that some of Sterne's sermons stand out above all the rest. I therefore propose to test this intuition/assumption through a comparison between Sterne's homiletic discourse and a corpus of contemporary sermons, in order to assess whether it is possible to reconcile the viewpoints of New and Goring.

My working hypothesis is that Sterne as a preacher may have dealt with typical homiletic ideas in a very original (hence 'eccentric') idiolect. His innovation may be more stylistic than doctrinal. Besides, eccentricity seems to be the characteristic feature of only a relatively small number of Sterne's sermons, which strike the reader as being more narrative, imaginative or even novelistic texts than standard post-Restoration pulpit oratory. Most of his other homilies sound much more conventional. Whether this is due to extensive plagiarism -- as first systematically analysed by Lansing Van der Heyden

Hammond -- or to typically Latitudinarian theology, as argued by New, remains to be seen.

Sterne's printed sermons will be compared with a full-text corpus of eighteenth-century English sermons, comprising works by Jonathan Swift, John Wesley, perhaps George Whitefield, the manuscripts of John Sharp (1723-92), and a subset of political sermons published in the first two decades of the century. Furthermore, internal comparisons between the collections of sermons which Sterne himself prepared for publication after the first instalment of *Tristram Shandy* and the three posthumous volumes published by his daughter are also necessary.

This paper will be based on the approach developed by the predominantly French school of lexicometry and stylometry, which emphasizes the use of exploratory multivariate statistical methods such as correspondence analysis and cluster analysis. For lack of a single software program that would ideally carry out all the necessary tasks, several packages will be used, especially *Hyperbase, Lexico3, Weblex, Wordmapper,* maybe *Wordsmith*. The linguistic and stylistic features identified by Douglas Biber as underlying different text types, especially the dimensions labelled by him as "narrative versus non-narrative concerns," "informational versus involved production," and "persuasion" will be explored.

References

Bandry-Scubbi, Anne & Deconinck-Brossard, F. (2005). 'De la lexicométrie à la stylostatistique : peut-on mesurer le style ?' *Bulletin de la Société de Stylistique anglaise* 26 *:* 67-85.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: CUP.

Coulthard, M. (2004). 'Author Identification, Idiolect, and Linguistic Uniqueness,' *Applied Linguistics* 25/4: 431-447.

Deconinck-Brossard, F. (1984). *Vie politique, sociale et religieuse en Grande-Bretagne d'après les sermons prêchés ou publiés dans le Nord de l'Angleterre, 1738 –1760*. Paris : Didier-Erudition.

---------- (1993). 'Eighteenth-Century Sermons and the Age,' in *Crown and Mitre : Religion and Society in Northern Europe since the Reformation*, eds W.M. Jacob & Nigel Yates. Woodbridge : The Boydell P, 105-21.

Delcourt, C. (2002). 'Stylometry,' *Revue belge de philologie et d'histoire,* 80/3 : 979-1002.

Downey, J. (1978). 'The Sermons of Mr Yorick: A Reassessment of Hammond,' *English Studies in Canada* 4/2: 193-211.

Goring, P. (2002). 'Thomas Weales's *The Christian Orator Delineated* (1778) and the Early Reception of Sterne's Sermons,' *The Shandean* 13: 87-97.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic.

Hammond, L. (1948). *Laurence Sterne's Sermons of Mr. Yorick*. New Haven : Yale U P.

Hawley, J. (1998). 'Yorick in the Pulpit,' *Essays in Criticism* 48/1: 80-88.

Kraft, E. (1996). 'A Theologic [sic] Flap Upon the Heart: *The Sermons of Mr. Yorick,'* in *Laurence Sterne Revisited*. New York: Twayne Publishers. Pp. 24-46.

Lebart, L et al. (1998). *Exploring Textual Data*. Boston; London: Kluwer Academic.

New, M. (1996). *The Sermons of Laurence Sterne: The Notes*. Gainsville: U P of Florida.

New, M., ed. (1996). *The Sermons of Laurence Sterne: The Text*. Gainsville: U P of Florida.

Deux romans baroques français en ligne

(L'Astrée d'Honoré d'Urfé et Artamène ou le Grand Cyrus de Madeleine et Georges de Scudéry) :

DÉMARCHE PHILOLOGIQUE ET PRATIQUES DE LECTURE

Delphine DENIS
Université Paris-Sorbonne(Paris IV)

Claude BOURQUI
Alexandre GEFEN
Université de Neuchâtel

Cette communication à trois voix se propose de mettre en perspective des projets éditoriaux consacrés à **deux longs romans français du XVII^e^ siècle** (*L'Astrée* d'H. d'Urfé, 1607-1627 et *Artamène ou Le Grand Cyrus* de G. et M. de Scudéry, 1649-1653), pour lesquels a été retenue la solution d'une **mise en ligne au format de la Text Encoding Initative, conçue comme un outil philologique et pensé dans un rapport de complémentarité avec une édition traditionnelle sur papier**. Il s'agira à la fois de présenter la démarche qui préside, selon des modalités différentes, à ces deux options, et de réfléchir sur le plan théorique aux diverses pratiques de lecture qu'engagent le couplage.

Le rapprochement de ces deux romans n'est pas motivé par la seule actualité scientifique, au demeurant inégale (le projet « *Artamène* » a déjà trouvé sa **réalisation concrète** sous http://www.artamene.org); le chantier de *L'Astrée* n'est ouvert que depuis un an environ). On rappellera qu'outre leur appartenance à une même période historique, un lien de filiation objective les unit, puisque le modèle du roman urféen est explicitement revendiqué par les auteurs du *Grand Cyrus*. D'autre part, un même postulat d'***illisibilité*** pèse sur leur lecture actuelle : trop longs, trop complexes dans leur structure narrative et énonciative, trop accueillants à l'égard de sous-genres hétérogènes (lettres, poésies, etc.), ces textes sont réputés, l'un aussi bien que l'autre, inaccessibles au lecteur moderne. On le montrera cependant, si ces deux textes relèvent de questionnements comparables, ils imposent cependant des **choix éditoriaux spécifiques** de traitement et de visualisation des textes sources encodés au format XML/TEI.

Or, en procurant des versions en ligne du *Grand Cyrus* et de *L'Astrée*, on ne modifie pas seulement de manière radicale les conditions d'accessibilité de ces deux romans s'étendant sur des milliers de pages. Cette approche favorise également une saisie des œuvres par « morceaux choisis », qui présente de grandes **analogies avec les pratiques de lecture originelles**, privilégiant les circulations « sélectives » à travers les textes – ainsi que l'atteste, entre autres, l'existence de « Tables » dans les éditions du XVII^e^ siècle (table des histoires insérées, des lettres, des poésies, des oracles etc.).

L'affinité structurelle possible de ces romans avec le support informatique en ligne rend donc particulièrement pertinente la création et le développement de sites procurant une édition intégrale, munie de **dispositifs de lectures parfois inédits** (résumés dépliables, fiches personnages, encyclopédie hypertextuelle, etc.). Mais les deux textes peuvent également retirer d'autres bénéfices majeurs de la migration sur ce nouveau support : dans la mesure, d'abord, où l'encodage en TEI des différentes versions du texte s'inscrit dans le cadre d'une **réflexion cruciale sur la genèse de l'œuvre** ; ensuite, parce que les **possibilités multimédia** offertes par l'édition électronique permettent de faire écho au « dialogue des arts » orchestré par ces deux romans (musique, arts décoratifs, architecture, iconographie, etc).

Mais la communication proposée ne se contentera pas de mettre en évidence ces convergences fondatrices. Elle s'attachera également à soulever quelques points de difficultés rencontrés et à commenter les diverses solutions qui y ont été apportées grâce à l'usage de la norme TEI :

- **la complémentarité entre livre imprimé et site Internet** : rapport anthologique ou double version intégrale ? visibilité prioritairement confiée à l'un des deux *media* ou aux deux ? exigences de rigueur philologique appliquées aux deux versions du texte ou à une seule ?
- **la détermination de l'usager privilégié du site** : lecteur d'agrément, chercheur en quête de données,

enseignant à la recherche d'un matériau pédagogique, voire les trois, grâce à un mode d'affichage du texte librement choisi en fonction des objectifs de chaque visiteur ?

- **l'alliance entre l'exigence de pérennité et la multiplicité des usages** : à la nécessité de figement et de conservation patrimoniale d'un texte s'oppose sa nécessaire adaptation à des modes de diffusion et de lecture hétérogènes (version word, version ebook, extraction *via* un moteur de recherche, exportation vers d'autres outils d'analyse textuelle, etc.).
- **la création d'une interface de lecture adaptée à un texte spécifique** mais compatible aussi bien avec les pratiques philologiques traditionnelles qu'avec les normes explicites et implicites de la navigation sur Internet, de la visibilité/indexabilité des pages ainsi produites, dans le strict respect d'un standard universel, la TEI.

Two French Baroque Novels on Line

(***L'Astrée*** by Honoré d'Urfé
and ***Le Grand Cyrus*** by Madeleine
and Georges de Scudéry):

PHILOLOGICAL PROCESS
AND READING PROCEDURES

In this paper, we would like to display editorial projects dedicated to **two long seventeenth-century French novels** (*L'Astrée* by H. d'Urfé, 1607-1627 and *Artamène ou le Grand Cyrus* by G. and M. de Scudéry, 1649-1653) which **have been, or will be, put on-line in Text Encoding Initiative format**. **This format was thought up as a philological tool and designed to be complementary with a traditional paper edition.** We would like to present the process used for these two options and its different methods and consider from a theoretical point of view the different reading practices that this coupling brings on.

It is not only the present scientific situation that brought us to draw a parallel between the two novels (the "*Artamène*" project already exists **in concrete terms** at http://www.artamene.org, whereas the "*Astrée*" site was undertaken only a year ago). As well as being from the same historical period, the two texts are linked by an objective relation, seeing that the authors of the *Grand Cyrus* explicitly claim they followed the model of d'Urfé's novel. Also, nowadays, they both suffer from the same postulate of being **unreadable**: too long, based on a narrative and enunciative structure that is too complex, including too many different subgenres (letters, poems, etc.), these texts both have the reputation of being inaccessible for the modern-day reader. However, as we will demonstrate, even though both texts raise the same questions, they each demand **specific editorial choices** for the processing and the visualisation of the source texts encoded in XML/TEI format.

By producing on-line versions of the *Grand Cyrus* and *L'Astrée*, not only do we radically change the accessibility of these two novels that stretch out over thousands of pages, we also encourage the website visitors to read "selected extracts" **in a way that the original reader would have**, favouring a selective path through the texts, as it is attested by the tables found in the seventeenth-century editions (tables of inserted stories, letters, poems, oracles, etc.).

The possible structural affinities of these novels with the on-line medium makes it then very relevant to create and develop websites offering an integral edition with **reading devices** ("unfoldable" summaries, character sheets, a hypertextual encyclopaedia, etc.) But for both texts, the migration to a new medium has its advantages: firstly, insofar as the TEI encoding of the text is done in the perspective of a **crucial consideration of the genesis of the text**; secondly, because the **multimedia possibilities** produced by an electronic edition allow us to echo the "dialogue of arts" created by the two novels (music, decorative arts, architecture, iconography, etc.)

However, in this paper, we will not only point out these different convergences, but we will also bring up some of the difficulties we came up against and comment on the different solutions that were found thanks to the TEI standards:

- **the complementarity between the printed book and the website**: should one of the two be an

anthology, or should both be an integral version? Should one of the two media be favoured? Should strict philological requirements be applied to both versions?

- **establishing who should be the favoured website-user**: a pleasure-reader, a scholar looking for information, a teacher in search of educational material, or all three, thanks to a visualisation of the text chosen depending on what each visitor is aiming at?
- **combining the requirements of durability and the multiplicity of uses**: the necessity of freezing and conserving a text as a patrimony is opposed to its necessity to adapt to different kinds of distribution and reading (Microsoft Word version, "e-book" version, extraction *via* a search engine, exportation to other textual analysis tools, etc.)
- **creating a reading interface adapted to a specific text**, but compatible as much with traditional philological procedures as with the explicit and implicit criteria of Internet navigating and visualising and indexing the pages produced, while complying with a universal standard, TEI.

The Exhibition Problem. A Real Life Example with a Suggested Solution

Øyvind EIDE

Unit for Digital Documentation at the Faculty of Arts, University of Oslo

Background

In a paper presented at ACH/ALLC 2005, Allen H. Renear et.al. describe a problem of potentially great significance (Renear 2005). They argue that:

> "In ordinary linguistic communication we often use a name to refer to something in order to then go on to attribute some property to that thing. However when we do this we do not naturally construe our linguistic behavior as being at the same time an assertion that the thing in question has that name. (Ibid, p. 176)"

Further, they claim that this distinction is over-looked when conceptual models based on encoded texts are developed.

In our work at the Unit for Digital Documentation at the University of Oslo, we have used XML encoded material as sources for several of our databases (Holmen 1996, Holmen forthcoming). The way this is done is by marking up texts both descriptively and interpretatively, followed by the use of software to extract information which is included in the databases. If Renear's argument is correct, we may infer that the databases include assertions which are based on information in the source texts that is, strictly speaking, not grounded in these texts. For example, we could be using a text as the source of a naming in the database while the naming is merely exhibited, and not asserted, in the text.

The false resolutions

Renear et.al. propose three possible resolutions to this problem, but they also state that all of these are false. Their resolutions are the following:

1. TEI encoding represents features of the text only.

2. The use of two arcs, i.e. "The Semantic Web community solution", which will be discussed below.

3. Exhibition is a special case of presupposition.

Based on the description of our work above, it should be obvious that resolution no. 1 is not an alternative for us. Semantic modelling of the real world on the basis of descriptions in texts is part of our work.

I find it difficult to understand how resolution no. 3 may represent a possible solution. Whether exhibition is a type of presupposition or not does not change the basic problem; i.e. in our case, the use of a text as the source of a naming which is merely exhibited in the text. The problem remains the same if the naming is also presupposed in the text, as long as it is not asserted.

I claim that resolution no. 2 is not false after all, and below I will demonstrate how the Conceptual Reference Model (CIDOC-CRM) will solve a similar problem in my example text. The CIDOC-CRM is a ontology developed in the museum community to be used for cultural heritage documentation.

My example text

In this paper, no general solution to the problems identified above will be proposed. However, I believe that the special solution that I propose could easily be generalized.

The text used in my example is based on the work of Major Peter Schnitler. In the 1740s, Major Schnitler was appointed by the Danish government to explore the border area between the northern parts of Norway and Sweden/ Finland. Significant parts of the text in the manuscript that he handed over to the Danish government consist of transcripts of local court interviews which were carried out by Schnitler in order to gather information about the local population as well as what they had to say about the border areas. The material includes information directly relevant to the border question, as well as general information of the areas in question, which corresponds to similar material collected through work carried out in Europe at the time (Burke 2000, pp. 128 f.).

The text fragments below are taken from the very first meeting described in the text (English translation from Danish by me):

[1] Of the Witnesses, supposed to be the most Cunning on the border issue, Were and stood up in the court 1: Ole Larsen *Riise*.

[2] For these the Kingly order was read out loud [...] and they gave their Bodily Oath

[3] Question: 1: What his name is? *Answer: Ole Larsen Riise* (Schnitler 1962, p. 1)

In these quotes, we find that several facts are asserted by the text.

Excerpt 1 claims the existence of a witness. We will call this witness x. Being a witness implies being a person. Thus, x is a person. We may also note that x is referred to by using the name "Ole Larsen *Riise*.", abbreviated "OLR" below.

Excerpt 2: Person x gave an oath to speak the truth.

Excerpt 3: Person x, according to the text, claims that his name is OLR. The source of the naming is person x, as spoken out loud at a specified place at a specified date in 1742. The text puts forward an assertion by person x that he is named OLR.

Modelling the semantic content from our perspective

My semantic model of these facts will include the following information:

Assertion	Source
1) There is an x who is a witness	The text
2) x is a person	The meaning of the word "witness" and "person" in this context
3) x gave an oath	The text
4) OLR is the name of person x	x

It is easy to describe the source of the three first assertions through CIDOC-CRM, by stating that they are documented in Schnitler's text:

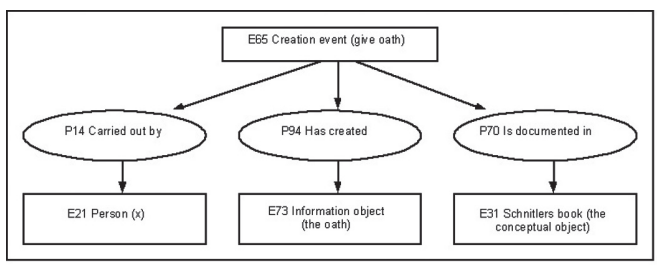


Figure 1

In this figure, as well as in the next one, the boxes with names starting with *E* represents entities, while the boxes with names starting with *P* represents the properties linking them together.

But how do we describe the source of the naming event? We start with the event in which the attribute was assigned (the naming event, a speech act), which is an *E13 Attribute assignment* which states that x carried out this particular speech act:

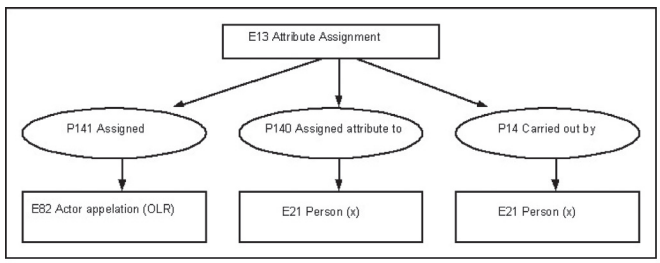


Figure 2

When looking at these two model figures, it is striking to what extent the modelling of the giving of the oath in Figure 1 compares to the naming of x in Figure 2. The explanation is that those are similar situations. Our traditional way of reading made us structure them differently in the table above, whereas represented in the CIDOC-CRM structure they came out the same in Figure 1 and 2. In order to show clearly in what way they correspond, note that line 4 in the table above could be rewritten as follows:

Assertion	Source
4) x named himself ORL	The text

This is a good example of the way modelling may help us understand a text better. What we have done is to rethink the difference between an event (x gave an oath) and a fact (ORL is the name of x). In order to model the fact correctly, i.e. to show that it was exhibited rather than asserted in the text, we had to consider it as a naming event. Considering it as an event is more feasible in that an event typically has actors who are responsible for the outcome. Further, this makes more sense in that both expressions are speech acts. When it is considered as a speech act, the naming event is the same kind of event as the giving of an oath.

Why solution 2 is not false after all

In order to be able to see the problem with Renear's solution no. 2, or to realize that the problem is not really there, we have to quote his text in extensio:

> "Another approach, this one anticipated from the Semantic Web community, is simply to insist on an unambiguous corrected conceptual representation: one arc for being named "Herman Melville", one for authoring *Moby Dick*. But this resolution fails for the reasons presented in the preceding section. Although this model would be in some sense an accurate representation of "how the world is" according to the document, it would not represent what is asserted by the document. The authorship arc in the corrected RDF graph model will correspond to relationships of exhibition, not assertion; and there is no accommodation for this distinction in the modelling language. (Renear, p. 178)"

In the first couple of sentences in this paragraph, the resolution of using an "unambiguous corrected conceptual representation" is said to have failed. The next couple of sentences weakens his statement by saying that only RDF does not accommodate this; "there is no accommodation for this distinction in ***the*** modelling language" (my emphasis). There are no arguments to support why a different modelling language could not solve the problem. In fact, the CIDOC-CRM does solve this, by giving the modeller an opportunity to state explicitly who is the source of an assertion, as demonstrated in Figure 2.

In the example above, we knew who made the assertion exhibited in the text. But even if we did not know, we could still make a similar model as long as we accept that it was made by somebody. In CIDOC-CRM, the modelling of entities we infer to exist without knowing who or what they are is quite possible.

Generalization

The example described above is quite special, as it includes an explicit naming. But it can be argued that all person names, at least in 18th century Scandinavia, are based on naming events, as people are baptised. As long as we believe that this is the case, we can include

in the model an explicit attribute assignment event as the one in Figure 2 for each name used in the text. This will be an event of which we do not know who carried it out or when it took place, but that is not necessarily a problem. The will always be things we do not know in historical texts. The naming event we model this way will also be an event that is not documented in the text we are basing the model on. Whether this is acceptable is a decision one has to take when building up such a model.

Conclusion

There is reason to believe that the problem described in Renear's paper is an important one. But a solution to the problem has been identified. I have shown that for one specific type of text, the problem may be solved by using CIDOC-CRM modelling including explicit statements of the sources of the assertions exhibited in the text. Further research may disclose whether this solution will work for other types of texts as well.

References

Burke, P. (2000) *A social history of knowledge : from Gutenberg to Diderot*. Cambridge.

CIDOC-CRM. *ISO/FDIS 21127. Information and documentation -- A reference ontology for the interchange of cultural heritage information* [Definition of the CIDOC Conceptual Reference Model].

Holmen, J.; Uleberg, E. (1996) "Getting the most out of it - SGML-encoding of archaeological texts." Paper at the IAAC'96 Iasi, Romania. URL: http://www.dokpro.uio.no/engelsk/text/getting_most_out_of_it.html (as of 2005-11-14).

Holmen, J.; Jordal, E.K.A; Olsen, S.A.; Ore, C.E. (forthcoming) "From XML encoded text to objects and events in a CRM compatible database. A case study". In: *Beyond the Artifact. Proceedings of CAA 2004, Computer Applications and Quantitative Methods in Archaeology.*

Parsons, T. (1990) *Events in the semantics of English : a study in subatomic semantics*. Cambridge, Mass.

Renear, A.H.; Lee, J.H.; Choi, Y.; Xiang, X. (2005) "Exhibition: A Problem for Conceptual Modeling in the Humanities". P. 176-179 in: *ACH / ALLC 2005. Conference Abstracts*. 2nd Edition, Victoria.

Schnitler, P. (1962) *Major Peter Schnitlers grenseeksaminasjonsprotokoller 1742-1745. Bind 1 [Major Peter Schnitlers border examination protocols 1742-45]* / by Kristian Nissen and Ingolf Kvamen. Oslo.

TEI, CIDOC-CRM and a Possible Interface between the Two

Øyvind EIDE
Christian-Emil ORE

Unit for Digital Documentation at the Faculty of Arts, University of Oslo

In the work of the TEI Ontologies SIG there have been an interest in finding practical ways of combining TEI encoded documents with CIDOC-CRM models. One way of doing so is including CIDOC-CRM information in a TEI document and linking CIDOC-CRM elements to TEI elements where appropriate. In this paper, this method is described through an example, together with an outline of the additional elements necessary in the TEI DTD used.

Background

In projects at the Unit for Digital Documentation, University of Oslo, we have created SGML and later XML encoded versions of printed and hand-written museum documents, such as acquisition catalogues, for more than ten years (Holmen 1996). To be able to store such documents in a standard format, we are planning to use TEI. Much of our material are archaeological documents, and there have been a growing interest in the use of XML in general and TEI in specific in archaeological community the last few years (Falkingham 2005, sec. 3.3, cf. sec. 4.3 and 5.2.3).

We also use CIDOC-CRM as a tool for modelling the contents of such tagged documents as they are read by museum professionals. We use this method to be able to include information from XML encoded documents in our museum inventory databases, with references back to the encoded documents (Holmen forthcoming). We would like to store CIDOC-CRM models in close relation to the TEI encoded document. This paper describes an example of how we try to define a syntax in which to store such datasets.

Extension of a TEI DTD

There are two different ways in which to extend a TEI DTD for inclusion of CIDOC-CRM models. We may include an element for each and every entity and property used in the model, or we may just include one TEI element for CIDOC-CRM entities and one for properties. We have chosen the latter method. This gives a limited and rather simple expansion of the DTD. This is similar to the way the XML version of the bibliographic standard MARC is designed (MARCXML).

This method will make it possible to create one document storing both textual markup and semantic interpretations of a text, while keeping the two parts of the document separate, except for links between specific elements in the two parts. This means that the document can be published as a text as well as form the base of an import to a database of records based on the interpretation, keeping the links back to the original text.

In this paper, we use a DTD fragment to show an outline of the extensions we need. The extensions is composed of a root crm element including a number of crmEntity elements and a number of crmProperty elements.

The root CIDOC-CRM element

```
<!ELEMENT   crm         (crmEntity*, crmProperty*)>
<!ATTLIST   crm
            id          #ID>
```

The entity element

```
<!ELEMENT   crmEntity   #PCDATA
<!ATTLIST   crmEntity
            id          #ID
            typeNumber  #NUMBER>
```

The property element

```
<!ELEMENT   crmProperty #EMPTY
<!ATTLIST   crmEntity
            id          #ID
            typeNumber  #NUMBER
            from        #IDREF
            to          #IDREF>
```

Example of use

A typical situation in which this approach could be used

is in archaeological documents. We have created a short dummy document containing some of the informations types commonly existing in our museum documents, as shown in Example 1.

The excavation in Wasteland in 2005 was performed by Dr. Diggey. He had the misfortune of breaking the beautiful sword in 30 pieces.

Example 1

A tagging of this could be made as in Example 2.

```
<p id="p1">The

    <ab id="e1">excavation in
     <name type="place" id="n1">Wasteland</name>
    </ab> in

    <date id="d1">2005</date>

    was performed by

    <name type="person" id="n2">Dr. Diggey</
    name>.

    He had the misfortune of

    <ab id="e2">breaking
     <ab id="o1">the beautiful sword</ab>
     in 30 pieces
    </ab>.

</p>
```

Example 2

There are many objects and relations of interest when modelling the archaeological world described in this text. A typical museum curator reading could include the elements shown in Table 1.

1. A place identified by a name documented in n1.
2. A person identified by a name documented by n2.
3. A time identified by a date documented in d1.
4. An event (the excavation) documented in e1.
5. An event (the breaking) documented in e2.
6. An object (sword) documented in o1.
7. Dr. Diggey participated in the excavation
8. Dr. Diggey and the sword participated in the breaking
9. The excavation took place at the place identified by a name documented in n1 and at a time identified by a date documented in d1.

Table 1

A possible CIDOC-CRM representation of one of the entities in Table 1, the excavation in line 4, is shown in Example 3. Included are also references to lines 2, 3, 7 and 9.

Note that Example 3 is only showing part of a model that would represent a normal archaeological reading of the paragraph above. E.g., the date should have a "is documented in" property such as the ones for the activity and the person, and the place (Wasteland) should be documented in a way similar to the person Dr. Diggey.

E7 Activity	--> P2 Has type	--> E55 Type[1]				
	--> P14 Carried out by	--> E21 Person	--> P131 Is identified by	--> E82 Actor appelation[2]	--> P70 Is documented in	--> E31 Document[3]
	--> P4 Has time-span	--> E52 Time-span	--> P78 Is identified by	--> E50 Date[4]		
	--> P70 Is documented in	--> E31 Document[5]				

1) archaeological excavation

2) Dr. Diggey

3) the element identified by the id "n2" in the text of Example 2 above

4) 2005

5) the element identified by the id "e1" in the text of Example 2 above

Example 3

Example 4 shows this using the TEI-CRM syntax outlined in the DTD addition above. The crm element holds the small CIDOC-CRM model we have expressed in a TEI syntax, while the link element holds connections between the CIDOC-CRM model and the TEI text from Example 2. In this example we see that although all the CIDOC-CRM information may be expressed in such a syntax, an XML validation of the document will only validate a part of the information. It will not check whether the model adheres to the rules for e.g. which CIDOC-CRM properties may be used in connection to

which entities.

```
<crm id="crm-mod1">
    <crmEntity id="ent1" typeNumber="7">
    </crmEntity>
    <crmEntity id="ent2" typeNumber="55">
    archaeological excavation</crmEntity>
    <crmEntity id="ent3" typeNumber="21">
    </crmEntity>
    <crmEntity id="ent4" typeNumber="82">Dr. Diggey
    </crmEntity>
    <crmEntity id="ent5" typeNumber="31">
    </crmEntity>
    <crmEntity id="ent6" typeNumber="52">
    </crmEntity>
    <crmEntity id="ent7" typeNumber="50">2005
    </crmEntity>
    <crmEntity id="ent8" typeNumber="31">
    </crmEntity>
    <crmProperty id="prop1" typeNumber="2"
    from="ent1" to="ent2"/>
    <crmProperty id="prop2" typeNumber="14"
    from="ent1" to="ent3"/>
    <crmProperty id="prop3" typeNumber="131"
    from="ent3" to="ent4"/>
    <crmProperty id="prop4" typeNumber="70"
    from="ent4" to="ent5"/>
    <crmProperty id="prop5" typeNumber="4"
    from="ent1" to="ent6"/>
    <crmProperty id="prop6" typeNumber="78"
    from="ent6" to="ent7"/>
    <crmProperty id="prop7" typeNumber="70"
    from="ent1" to="ent8"/>
</crm>

<linkGrp type="TEI-CRM interface">
    <link targets="#ent5 #n2"/>
    <link targets="#ent8 #e1"/>
</linkGrp>
```

Example 4

Conclusion and further research

While different uses of ontological models in connection to TEI documents will differ in their technical solutions, e.g. whether the ontological model rests in a separate document or not, and which syntax is chosen for the model, the three main elements shown here have to be present:

- a TEI document
- an ontological model expressed in some XML syntax
- link elements to connect the specific elements from the two together

We have described a way of expanding TEI that gives us the tools we need to include a CIDOC-CRM model in a TEI document, and connect specific CIDOC-CRM entities to specific TEI elements in the non-CRM part of the document. We would like to see research into similar methods of connecting informations in other ontological systems to TEI documents, to discover whether a similar method is feasible. It would also be interesting to see if it is possible to make a general addition to TEI for this use, or if each ontological system needs its own tag set.

In our own research, we will write out an ODD to test this method on samples of our own data, and then continue to implement this model on real data, so that the usability of this method for complete documents and CIDOC-CRM models can be examined.

References

CIDOC-CRM. ISO/FDIS 21127. Information and documentation -- A reference ontology for the interchange of cultural heritage information [Definition of the CIDOC Conceptual Reference Model].

Falkingham, G. (2005) A Whiter Shade of Grey: a new approach to archaeological grey literature using the XML version of the TEI Guidelines. Internet Archaeology, issue 17. URL: http://intarch.ac.uk/journal/issue17/falkingham_toc.html (as of 2005-11-14).

Holmen, J.; Uleberg, E. (1996) "Getting the most out of

it - SGML-encoding of archaeological texts." Paper at the IAAC'96 Iasi, Romania. URL: http://www.dokpro.uio.no/engelsk/text/getting_most_out_of_it.html (as of 2005-11-14).

Holmen, J.; Jordal, E.K.A; Olsen, S.A.; Ore, C.E. (forthcoming) "From XML encoded text to objects and events in a CRM compatible database. A case study". In: Beyond the Artifact. Proceedings of CAA 2004, Computer Applications and Quantitative Methods in Archaeology.

MARCXML. MARC 21 XML Schema. URL: http://www.loc.gov/standards/marcxml/ (as of 2005-11-14).

TEI Ontologies SIG. URL: http://www.tei-c.org/Activities/SIG/Ontologies/ http://www.tei-c.org/wiki/index.php/SIG:Ontologies (as of 2005-11-13).

TEI P5 (2005) Guidelines for Electronic Text Encoding and Interchange. [draft] Version 0.2.1. TEI Consortium, 2005.

Modelling Lexical Entries: a Database for the Analysis of the Language of Young People

Fabrizio FRANCESCHINI
Elena PIERAZZO

University of Pisa - Department of Italian Studies

At the Humanities Faculty of University of Pisa, we started a big project devoted to the study of the language and culture of the young people. The enquiries were initially held in the area of provenance of students of the University of Pisa, including the districts of Massa-Carrara, Lucca, Pistoia, Pisa, Livorno, Grosseto and the district of La Spezia in the Liguria region. We distributed a questionnaire among students of the last two years of secondary school (around 18 years old). The enquiries took place in towns that host secondary schools, i.e. towns and big villages of the urbanized country.

The questionnaire includes:

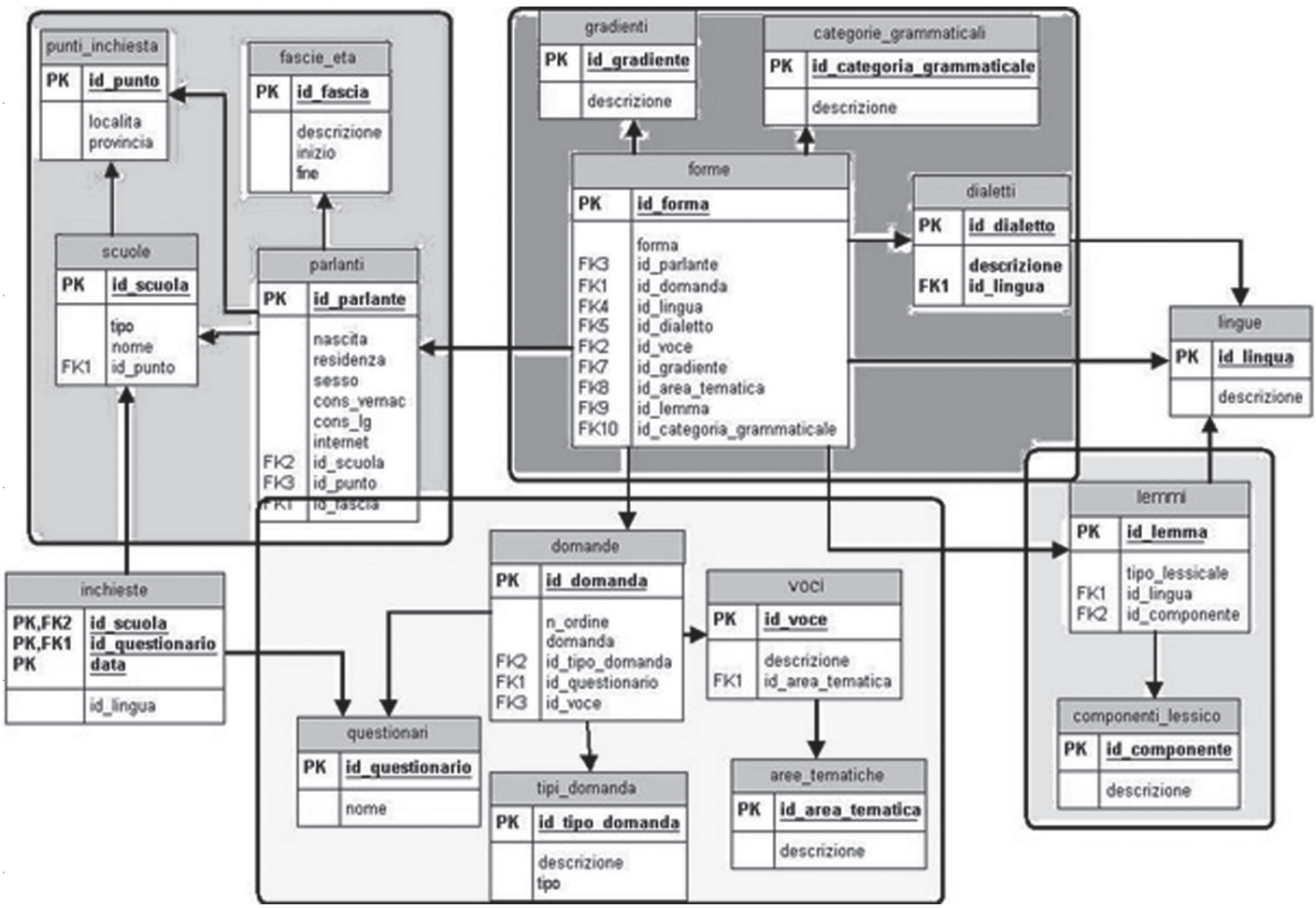
1. a socio-linguistic section enquiring on the social condition, cultural preferences and style of life; usage of the dialect inside the family and usage of forms exclusive for young people;
2. a lexical section that counts 36 onomasiologic questions ("How do you say for…") referred to the different spheres of the life of young people (family, school, external world, interpersonal relations, judgements, etc.)
3. open fields for spontaneous linguistics insertions.

The enquiries involved 2.500 informants and produced over 70.000 forms. Because of this huge mass of data we have looked for a suitable storage solution that would have let us easily query the data. First of all, we wanted to query the forms themselves in a quick and simple way. Secondly, we wanted to group the results according to different parameters such as place of enquiry, sex of the informants, and socio-cultural divisions. Others goals

were the possibility of measuring the dialectal or the mass-media language influence and focalizing any lines of development of the Italian language.

It was immediately clear that the results of the enquiries could not be directly reversed into a digital format because of a lack of structured information. Furthermore, we felt the need of analyzing and classifying the produced forms from a linguistic point of view, for example tracing every form to the relevant lemma and recognizing the grammar category in order to enable sophisticated queries.

Firstly we tried to mark data in the XML language, using the TEI Terminological Databases tag set, but this attempt showed some limitations from the very beginning. As we know, XML works at its best with semi-structured data such as texts, on the contrary our lexical entries are strongly structured data, and are connected with both linguistic information and personal data supplied by informants. Performing such links through XPath or ID-IDREFS system proved to be quite a farraginous mechanisms. Furthermore the TEI Terminological Database tag set offers just generic elements, providing poor description of lexical entries; the TEI Consortium itself has considered unsatisfactory such tag set that will be strongly revised in the forthcoming P5 version (see http://www.tei-c.org/P5/Guidelines/TE.html). In addition, reversing in XML the data pertinent to two points of enquiry, we obtained so large files that most common XML applications showed remarkable difficulties in managing them. For all these reasons, we decided to reverse the data in a MySQL relational database that would have let us to bypass such limitations. A relational DB is certainly a more suitable and performing solution for strongly structured data as we conceived our entries.



[FIG. 1] BaDaLì conceptual diagram

We called our database BaDaLì, acrostic for *Banca Dati Linguagiovanile*, but also an expression that means 'look over there' or 'mind you' or even, in an antiphrastic sense very common in young people language, 'never mind'. *Figure 1* shows a simplified diagram of the BaDaLì database.

BaDaLì can be ideally subdivided in four main modules: the informants module (green area), the questionnaires module (yellow area), the lemmas module (blue area) and the forms module (orange area); table *lingue* ('languages') is a lookup shared by lemmas and forms modules, while table *inchieste* ('enquiries') connects the informants module with the questionnaires module.

Forms module

The very central point of the database is the table *forme* ('forms') and contains the forms produced by informants. Every form is traced to its grammar category (*categorie grammaticali* table). In some cases the form is also related to a specific dialect (*dialetti* table), or to a language (*lingue* table) in case of foreignism. In this way an eventual existing distance between form and lemma (that we call *gradiente* 'gradient') is measured in terms of dialectal influence, foreign features or innovative traits on a graphical level.

The forms module is connected to all other modules. Every form, in fact, is traced to its relevant lemma, is produced by an informant, under the stimulation of a question.

Lemmas module

Tracing forms to their relevant lemmas is a crucial point, especially for innovative forms not recorded by dictionaries. For this reason, we selected a reference dictionary and established a number of criterions to create the suitable lemma for the unattested occurrences. We decided to adopt the most complete dictionary of modern Italian, the *Grande Dizionario Italiano dell'Uso* (Gradit), edited by Tullio De Mauro. When a new lemma is inserted in the database, specific codes are added to mark its absence or a semantic innovation in regards of Gradit.

Relevant lemmas (*lemma* table) have been categorized too, following an updated version of the classification of lexical components of the language of young people (*componenti lessico* table) proposed by Cortelazzo 1994.

Informant module

Table *parlanti* ('informants') collects all the information pertinent to the informants. A number of questions pointed out the need of typifying data collected by the enquiry. For example, the questionnaire asked the informer to declare his/her birthplace and residence. The result was a list of towns and villages that had little relevance in the case of a large scale enquiry. As the question about birthplace was included to retrieve information about the origin of the informant's family in order to determine if an influence of a non local dialect can be assumed, we decide to group answers in macro-regional categories. The question on the residence was introduced to retrieve information about the commuting, to enquire on eventual differences between the language of towns and small villages. As the secondary schools where the enquiries were held are located in towns or in big villages, we decided to consider only if the informant lives in the same town where the school is located or elsewhere. In such cases a relative loss of information is compensated for the opportunity of comparing the results of different points of enquiry.

Questionnaires module

Questionnaires module collects data about questionnaires and questions. We took into account the possibility of inserting lexical entries produced by different questionnaires or by updating of our questionnaire. Therefore the term enquired by the questions (*voce indagata* table) is isolated. For example, in case of the question "How do you say for money?", the word "money" is the enquired term; the comparison of forms produced by different questionnaires enquiring on the same term is easier by isolating the enquired term form the body of the question.

Our database includes onomasiologic questions ('How do you say for…') and thematic questions ('Which words do you know about…'), so a classification for different types of questions (*tipi domanda* table) was needed.

BaDaLì public interface

The database is currently freely available on the Web at the address http://dblg.humnet.unipi.it.

[FIG. 2] BaDaLì home page

The provided interface allows a number of possible queries:

1. starting from a **lemma** (*Tipi lessicali*), it is possible to retrieve all the forms traced to such a lemma; a further step allows grouping the results according to three parameters: place of the enquiry, sex, and kind of school.
2. starting from the **enquired term** (*Voci Indagate*), it is possible to retrieve all the forms produced under the stimulation of such a term. A further step allows grouping the results according to the same three parameters as in 1.
3. starting from a **form** (*Forme*), it is possible to retrieve forms grouped according to the three parameters as in 1.

Modelling is certainly a crucial point in designing new projects, since the design will determine from the very beginning which requests a tool will be able to satisfy. In this frame we propose our experience, hoping to stimulate a reflection on such a topic.

References

Cortelazzo, M. (1994). Il parlato giovanile. In: *Storia della lingua italiana*, vol. 2, *Scritto e parlato*, ed. by Luca Serianni /Pietro Trifone, Torino, p. 291-317.

De Mauro, T. (1999). *Grande Dizionario Italiano dell'Uso*. (Gradit). Torino.

Sperberg-Mcqueen, C.M. and Burnard, L. (2002). *TEI P4 Guidelines for Electronic Text Encoding and Interchange: XML-compatible Edition*. (P4). Oxford, Providence, Charlottesville, & Bergen. Available also at http://www.tei-c.org/P4X/.

Collaborative Indexing of Cultural Resources: Some Outstanding Issues

Jonathan FURNER
OCLC Online Computer Library Center, Inc., Dublin, OH, USA

Martha SMITH
Information School, University of Washington, Seattle, WA, USA

Megan WINGET
School of Information and Library Science, University of North Carolina at Chapel Hill, NC, USA

In this paper, we argue that, although collaborative indexing of cultural resources shows promise as a method of improving the quality of people's interactions with those resources, several important questions about the level and nature of the warrant for basing access systems on collaborative indexing are yet to receive full consideration by researchers in cultural informatics. Specifically, we suggest that system designers look to three cognate fields---classification research, iconographical analysis, and annotation studies---for pointers to criteria that may be useful in any serious estimation of the potential value of collaborative indexing to cultural institutions.

Collaborative indexing (CI) is the process by which the resources in a collection are indexed by multiple people over an ongoing period, with the potential result that any given resource will come to be represented by a set of descriptors that have been generated by different people. One community of researchers that has demonstrated heightened, ongoing interest in collaborative indexing is that which is active in *cultural informatics*, and specifically in the design and development of systems that provide patrons of cultural institutions such as libraries, archives, and museums with networked access to digital representations of the objects in institutions' collections (see, e.g., Bearman & Trant, 2005).

Justifying the collaborative-indexing approach

At a simple level, the quality of any cultural information system (or any component of a system such as an indexing mechanism) may be evaluated by considering its performance in relation to three imperatives, each of which corresponds to a separate aspect---cultural, political, economic---of the complex mission of contemporary cultural institutions.

1. How *effectively* does the system allow its users to find the resources in which they have an interest, and to derive optimal value from those resources once found? In order that patrons derive positive value from their experience of interacting with the objects preserved in institutions' collections, they should be actively supported in their efforts to develop an interpretive understanding of those objects and the contexts in which they were produced---both by being given high-quality access to information about (including visual images of) objects, and by being encouraged to express their understanding and share their interpretations with others.
2. How broadly and *inclusively* does the system serve all sections of its parent institution's public? Managers of cultural institutions commonly express a concern that the opportunity to derive positive value from the services offered by those institutions should be distributed justly among, and enjoyed equally by, members of all social groups.
3. How well does it do at delivering maximal quality at *minimal cost*? The institution which consistently allows the costs incurred in the collection, preservation, interpretation, and provision of access to its resources to exceed the value of the benefits enjoyed by its public will not survive for long.

Justifications of the CI approach tend to proceed by drawing attention to the ways in which it can be viewed as responding to one or other of these three imperatives. Proponents commonly highlight several distinctive characteristics of CI in this regard:

(a) CI is *distributed*: No single person is required to index all resources; no single resource needs to be indexed by all people.

(b) CI is *cheap*: Indexers typically volunteer their

efforts at no or low cost to collection managers.

(c) CI is *democratic*: Indexers are not selected for their expertise by collection managers, but are self-selected according to indexers' own interests and goals.

(d) CI is *empowering*: People who might in the past have been accustomed to searching databases by attempting to predict the descriptors used by "experts" are now given the opportunity to record their own knowledge about resources.

(e) CI is *collaborative*: Any given record is potentially representative of the work of multiple people.

(f) CI is *dynamic*: The description of a given resource may change over time, as different people come to make their own judgments of its nature and importance.

All of these characteristics are relevant, in various combinations and to various degrees, to any estimation of the success with which CI-based systems are likely to meet the cultural, political, and economic imperatives described above. But each additionally raises issues of a more problematic nature than is typically admitted. Given the distributed nature of CI, for example, how can it be ensured that every resource attracts a "critical mass" of index terms, rather than just the potentially-quirky choices of a small number of volunteers? Given the self-selection of indexers, how can it be ensured that they are motivated to supply terms that they would expect other searchers to use? Empirical, comparative testing of the utility of different prototypes---focusing, for example, on forms of interface for elicitation of terms, or on algorithms for the ranking of resources---is undoubtedly an essential prerequisite for the future development of successful CI-based systems (Bearman & Trant, 2005). But it is also important, we argue, that the results of prior research in a variety of cognate fields be taken into account when addressing some of the more problematic issues that we have identified.

Classification research

In *classification research*, for example, it has long been argued that indexers and searchers benefit from having the opportunity to *browse* or navigate for the terms or class labels that correspond most closely to the concepts they have in mind, rather than being required to specify terms from memory (see, e.g., Svenonius, 2000). Indexer--searcher consistency, and thus retrieval effectiveness, can be improved to the extent that a system allows indexers and searchers to identify descriptors by making selections from a display of the descriptors that are available to them, categorized by facet or field, and arranged in a hierarchy of broader and narrower terms so that the user can converge on the terms that they judge to be of the most appropriate level of specificity. Current implementations of CI-based systems shy away from imposing the kind of vocabulary control on which classification schemes and thesauri are conventionally founded: the justification usually proceeds along the lines that indexers should be free, as far as possible, to supply precisely those terms that they believe will be useful to searchers in the future, whether or not those terms have proven useful in the past. Yet it remains an open question as to whether the advantages potentially to be gained from allowing indexers free rein in the choice of terms outweigh those that are known to be obtainable by imposing some form of vocabulary and authority control, by offering browsing-based interfaces to vocabularies, by establishing and complying with policies for the specificity and exhaustivity of indexing, and by other devices that are designed to improve indexer--searcher consistency.

Theories of iconographical interpretation

Another related subfield of library and information science is that which is concerned with the effective provision of subject access to art images (see, e.g., Layne, 1994), and commonly invokes the theory of iconographical interpretation developed by the art historian Erwin Panofsky (Panofsky, 1955). Current implementations of CI-based systems for art museums focus on eliciting generic terms for (what Panofsky calls) *pre-iconographic* elements, i.e., pictured objects, events, locations, people, and simple emotions---the assumption apparently being made that such terms are those that will be most useful to searchers (Jörgensen, 2003). There is very little evidence supplied by studies of the *use* of art image retrieval systems, however, to suggest either that pre-iconographic elements are indeed what non-specialists typically search for, or that generic terms lead non-specialist searchers to what they want. We do know from analyses of questions that visitors ask in museums that non-specialists typically do not have the specialist

vocabulary to specify precisely what they are looking for (see, e.g., Sledge, 1995). This does not necessarily mean, however, that searchers always default to using pre-iconographic terms whenever they wish to get at more complex themes and ideas, nor that searches for higher-level elements using pre-iconographic terms will be successful. Further studies of the question-formulating and searching behavior of non-specialist art viewers and learners are clearly necessary.

Annotation studies

Researchers in the human--computer interaction (HCI) community are continuing to develop an agenda for work in the emerging subfield of *annotation studies* (see, e.g., Marshall, 1998), focusing on ways to improve interfaces that support annotation behavior of a variety of kinds, in a variety of domains. In this research, an annotation is commonly considered as evidence of a reader's personal, interpretive engagement with a primary document---a form of engagement that is not so different from that which cultural institutions seek to encourage in their patrons. A cultural annotation system that allowed patrons not only to supply their own descriptions of an institution's resources, but also to add comments and to build communities around personal collections, could be envisaged as a vital service that would help patrons interact with and interpret those resources, largely outside the authority and control of curators and other specialists. It remains an open question as to whether a system that allows patrons to supply their own descriptions of institutions' resources is most appropriately evaluated as a tool for creating and accessing personal annotations, as a tool for sharing and accessing collaborative descriptions, as a retrieval tool pure and simple, or some combination of all three. Unfortunately, our understanding of the purposes and intentions of users of CI-based systems is still spotty, and further research in this area is necessary.

Conclusion

In general, we suggest that particular care needs to be taken by cultural institutions in examining and adjudicating between potentially conflicting motives for inviting patrons to provide basic-level descriptions of resources. Classification research shows us that simple assignment of single-word descriptors unsupported by vocabulary control or browsable displays of the semantic relationships among descriptors is not enough to guarantee effective access; theories of iconographical interpretation demonstrate how important it is that non-specialist indexers should not be led to assume that listing what one sees is somehow all that art-viewing and meaning-making involves; and annotation studies encourage us to consider how cultural institutions may go beyond simple systems for collaborative description, and develop more-sophisticated systems for truly collaborative annotation that support deeper levels of interpretation and learning.

References

Bearman, D. and Trant, J. (2005). Social terminology enhancement through vernacular engagement: Exploring collaborative annotation to encourage interaction with museum collections. *D-Lib Magazine*, **11**(9). http://www.dlib.org/dlib/september05/bearman/09bearman.html (accessed 7 November 2005).

Jörgensen, C. (2003). *Image Retrieval: Theory and Research*. Lanham, MD: Scarecrow Press, 2003.

Layne, S. S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science*, **45**: 583-588.

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In *Proceedings of ACM Hypertext '98* (Pittsburgh, PA, June 20-24, 1998), pp. 40-49. New York: ACM Press.

Panofsky, E. (1955). Iconography and iconology: An introduction to the study of Renaissance art. In *Meaning in the Visual Arts*. New York: Doubleday.

Sledge, J. (1995). Points of view. In D. Bearman (ed), *Multimedia Computing and Museums: Selected Papers from the Third International Conference on Hypermedia and Interactivity in Museums* (ICHIM '95 / MCN '95; San Diego, CA, October 9-13, 1995), pp. 335-346. Pittsburgh, PA: Archives & Museum Informatics.

Svenonius, E. (2000). *The Intellectual Foundation of Information Organization*. Cambridge, MA: MIT Press.

The Septuagint and the Possibilities of Humanities Computing: from Assistant to Collaborator

Juan GARCES

Centre for Computing in the Humanities

The Septuagint is in many ways a remarkable collection of texts. It represents the first known attempt to translate the Hebrew Bible into an Indo-European language, namely Hellenistic Greek. As such, it functions as an invaluable source for understanding pertinent linguistic, translational, text-critical, socio-cultural, and philosophical-theological issues that led to its creation and reception. Spanning in its inception from the first translations in the early- to mid-third century BCE to the later revisions in the second century CE, it gives scholars an insight not only into the development of the Greek language, but also into the influence of a Semitic language on its vocabulary and possibly even its syntax. Furthermore, being one of the rare cases where both a translated Greek text and the Semitic source text are extant, it also offers a rich source of insight into contemporary translation techniques and philosophies, albeit influenced by its hagiographic status, and helps in establishing the possibility of a clearer understanding of other Greek texts that are generally deemed to be translations from Semitic originals. Last, but not least, is its reflection of the culture and ideology of diaspora communities in the eastern Mediterranean metropoles, which led to the emergence and shaping of two important religious groupings.

The Septuagint is also amongst the ancient texts to receive early concerted applications of Humanities Computing approaches. The most prominent project in this line is the Computer Assisted Tools for Septuagint Studies (CATSS) project, which was called into life through the initiative of the International Organization for Septuagint and Cognate Studies in the early 1970s. Located at the University of Pennsylvania's emerging Center of Computer Analysis of Texts (CCAT), this project sought the use of computing resources towards three goals: (1) a morphological analysis of the Greek text, resulting in a tagged electronic text, (2) the comparison of Hebrew and Greek texts, resulting in an electronic parallel aligned Hebrew and Greek text, and the recording of published critical variants, resulting in an electronic Greek texts, with variants encoded. All these texts are now freely available (upon the signing of a user agreement/declaration) at the CCAT's gopher and form the basis of most, if not all, current Septuagint projects and studies making use of computing.

Departing from this pioneering, indispensable, and foundational application of Humanities Computing approaches to the study of the Septuagint, this paper will present an appraisal of the Humanities research questions presently asked of these texts, on the one hand, and of the potential of applying Humanities Computing in answering them, on the other. Beyond the widely agreed proposals of transforming such resources into established formats, e.g. Unicode for character encoding and TEI XML for text encoding, I will seek to discuss concrete problems and possibilities in pursuing Humanities Computing applications to the Septuagint, while generalising some of the insights within a wider context. The wider question will be: What should be done to and with electronic text(s) of the Septuagint in order to enrich it as a resource for answering the philological, historical, socio-cultural, and theological questions currently asked about it?

One of the aims of this paper will be to tease out the current disciplinary boundaries between traditional Humanities approaches and emerging Humanities Computing ones and to identify important developments in their relationship. An important presupposition in this discussion will be the understanding that Humanities Computing, as a hybrid discipline, will only be truly successful if it reflects a thorough understanding of both Humanities research questions and Computing approaches and develops a balanced negotiation of models and concepts that successfully bridge between the two. The direction of proposing research questions has to be pursuit in both directions – both 'how can one exploit Computing approaches to answer Humanities questions?' and 'How do Computing approaches alter the Humanities questions we ask about a research object?' have to be asked. To push further the metaphor in the name of the aforementioned CATSS project: It is a matter of using computing approaches as collaborators, rather than as mere assistants.

There are a number of issues in the case of the Septuagint that complicate straightforward conceptual models. To

choose but one illustration: Both the textual bases of Hebrew source text and Greek translation text are disturbed not only by the textual variants on each side, but also by the fact that the Septuagint texts soon encountered rival Greek translations, in some cases clearly influencing later revisions. Furthermore, the Hebrew source text for the Septuagint clearly departs occasionally from the Hebrew Masoretic Text. Moreover, the Septuagint also includes a number of apocryphal books, some of which were probably written directly in Greek. It is evident that the conceptual model to deal with this scenario cannot just consist of the juxtaposition of two clearly delimited texts. But how does one model such a complicated picture and how does one approach such a picture by computational means? This paper will attempt to propose some answers to this question. It will seek to do so by sketching an ontology and incipit model to accommodate the complication.

As an example of the wider context dealt with in this paper, I will discuss another, more general current development in Humanities research projects, not least influenced by contemporary communication technologies: the collaborative nature of the research undertaken. While the negotiation of consensus remains a crucial achievement and necessity in such endeavours, what are we to do with disagreement? How do we encode minority opinions and use them as a resource in computational approaches? If the underlying arguments are important: How do we record them for both agreements and disagreements?

Selected Literature

Septuagint

Jennifer Dines, *The Septuagint, Understanding the Bible and Its World*, Edinburgh: T. & T. Clark, 2004.

Natalio Fernández Marcos, *The Septuagint in Context: Introduction to the Greek Version of the Bible*, translated by W. G. E. Watson, Leiden: E. J. Brill, 2000.

CATSS

Computer Assisted Tools for Septuagint/Scriptural Study.

Robert A. Kraft, *"How I Met the Computer, and How it Changed my Life"*, SBL Forum (Spring 2004) .

Capturing Australian Indigenous Perceptions of the Landscape Using a Virtual Environment

Stef GARD
Sam BUCOLO
Theodor WYELD

School of Design, Queensland University of Technology Information Environments Program, ITEE, University of Queensland Australasian CRC for Interaction Design (ACID), Brisbane, Australia.

Introduction

The Digital Songlines project differs from the approach taken by most others in the field of virtual heritage. While there are many examples of recreated cultural sites, most of them are of a built form, such as temples, monuments, cities and townships. They are frequently re-created in 3-dimensions with a high level of realism. On the other hand, the Digital Songlines project's focus is on more than simple visualisation, rather its mission is to recreate an experience; a way of interacting with the simulated environment by identifying the key elements give to each place and its special cultural significance that an Aboriginal group identifies as being within their own tribal boundaries. Integrating the key cultural elements in a synthetic environment goes some way towards providing a setting for exploring otherwise inaccessible or previously destroyed significant sites. While traditional virtual heritage reconstructions frequently depend on technological solutions, Digital Songlines depends more on an understanding of the traditional cultural values attached to specific landscape by the participating Aboriginal people and then on a methodology and process for integrating those values in the digital environment with a focus on cultural relevance independent of its level of visual realism.

The continuing traditional culture of Australian Aborigines is one of the most ancient in the world. Recent research suggests that it is at least 40000 years old. European

colonisation of Australia since the late eighteenth-century, and farming, mining, tourism and social impacts of modern civilisation have since threatened this most remarkable cultural heritage. Aboriginal cultural custodians realise the urgent need to preserve the evidence of Australian Aboriginal heritage and culture to give young and future generations of Australian Aborigines a chance to identify with their aboriginal roots.

However, creating an Australian Indigenous cultural heritage environment causes some difficulties. Australian Aboriginal people perceive, in the landscape, details that non-indigenous people often fail to appreciate. Such details are very much a part of Aboriginal knowledge, spirituality and survival as well as cultural heritage. The landscape is perceived as a cultural entity, and needs to be recognised in a synthetic environment by Aboriginal people if the cultural heritage environment is to be perceived as authentic. One of the difficulties in undertaking such a task is the re-presentation of Aboriginal knowledge. There are few written records, hence it is mostly through the process of interviewing of cultural custodians that information can be gathered. This poses another problem: Aboriginal cultural custodians are not always comfortable with the traditional western research methods of interviewing and recording (AITSIS, 2000); they may prefer to tell their stories in a location which relates to the cultural context of the story, "Aboriginal reading comes out of the land, each place is a repository for information that is rarely commented upon elsewhere in the abstract but is released or stimulated by the place itself" (Strang, 2003, p200) (see figure 1). It is in this context that the Digital Songlines project has set itself the task of collecting cultural knowledge from Australian Aboriginal cultural custodians and Knowledge keepers, and, to provide a means of sharing that knowledge with future generations of Australian Aboriginal people through a virtual environment.

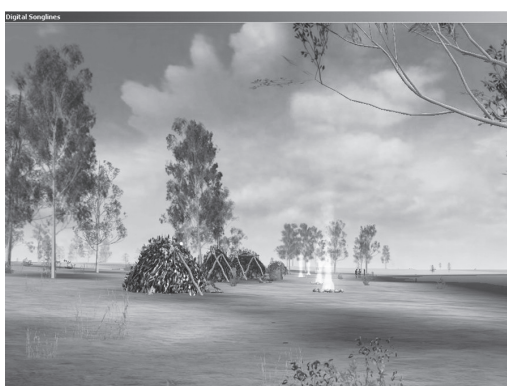


Figure 1. Real or simulated 'Country' is required contextualise to Australian Aboriginal story telling narratives.

Collecting Indigenous Cultural Knowledge

The need to 'locate' the telling of a story by an Aboriginal cultural custodian where access to the original environment is not possible involved reconstructing some locations using computer generated 3D models. This offers the advantage of portability and flexibility. However, the success of this method relies on its acceptance by the cultural custodians of the synthetic environment as a valid context for sharing their cultural knowledge. Two steps were implemented to try to achieve this. The first step was to identify the elements of the natural landscape that gave it a cultural meaning in the eyes of Indigenous people. The second step was to define a methodology to recreate these elements in a synthetic environment in a way that cultural custodians could recognise the cultural elements. Within the context of this paper the following describes the approach used to identify the elements of the landscape; it explains the protocols for approaching Australian Aboriginal people and a way of enlisting their trust.

Understanding the cultural elements of the landscape.

To-date some research with Australian Aboriginal people has resulted in suspicion and mistrust. As such , it is essential to inculcate participants in any study at every stage. The Cultural Custodians are the elders of their communities and not generally familiar with virtual reality and multi-media technologies. Therefore, it is necessary to demonstrate the potential of the technology. In this report we discuss an initial study that used a virtual environment showing landscape and approximately 10,000 year-old rock art from Mt Moffat in the Carnavon Gorges National Park in Central Queensland, Australia (see figure 2)**.** It was shown to a group of Australian Aborigines from central Queensland and their reactions observed.

As the initial goal was to gain the trust of the community, no recording or formal interview took place. Initially the community was suspicious of the researchers and of the technology but after three hours of community consultation, the community gained a better understanding of the technology and, most importantly, the intentions of the researchers.

Following the success of the initial contact, a cultural tour of the region was organised by two of the cultural custodians of the community. This allowed for more

observation even though there could still not be any formal recording of the event. One example of an observation related to the tradition of collecting food. Gathering bush food is a cultural activity. Australian Aboriginal people don't find food in the wild, they believe food is provided to them by Country (a Eurocentric analogy might be a form of benevolent land genie). However, there are conditions. Aboriginal people believe that one has to look after Country if one expects Country to provide for one. That is why, in their view, so many white people died of thirst and starvation in the early white settlement days, because they did not respect the Aboriginal Country law. Looking after Country means more than caring for it, it includes respecting the rituals taught by the ancestors

Figure 2. Approximately 10,000 year-old rock art from Mt Moffat in the Carnavon Gorges National Park in Central Queensland, Australia.

Converting the Observations to VR

The most important thing to come out of these observations was the attention to detail leading to a "contextual accuracy" that is very important to all the Aboriginal people encountered. Hence, in re-presenting a specific landscape in VR the following information should be gathered from the site. The importance of flora and fauna to a culturally recognisable landscape means more than realism alone. Realism has only a marginal effect on immersion in a virtual environment (Gower, 2003; Lee, 2004; Lombard and Ditton, 1997; Slater, et al, 1996). Hence, minimal realism yet culturally accurate virtual environments allows people without access to advanced technology, including Indigenous people in remote communities, to gain better access to the Digital Songlines Project and satisfies the needs of those communities for cultural sensitivity to representation of their stories about the land.

Conclusion

There are at least two issues facing the design of Aboriginal virtual heritage environments, contextual and cultural accuracy and realism. As the context is one of the most important aspects of culture sharing within Australian Aboriginal society then we need to know more about how to re-create a meaningful context in a virtual environment. It may be easier to evoke this by using reduced detail than highly realistic environments. In order to achieve this we need to learn from Aboriginal people what are the best 'signs' that can be used to identify their environments as unique and culturally significant.

Acknowledgment

This work is supported by ACID (the Australasian CRC for Interaction Design) established and supported under the Cooperative Research Centres Programme through the Australian Government's Department of Education, Science and Training.

References

Australian Institute of Aboriginal and Torres Strait Islander Studies (AITSIS), *Guidelines for Ethical Research in Indigenous Studies***.** 2000.

Strang, V. (1997). Uncommon ground: cultural landscapes and environmental values. Oxford: New York.

Gower, G. (2003). *Ethical Research in Indigenous Contexts and the practical implementation of it: Guidelines for ethical research versus the practice of research***.** Edith Cowan University, Perth, Western Australia.

Lee, K.M. (2004). *Presence explicated*. Communication Theory, p. 27-50.

Lombard, M. and Ditton, T. (1997). *At the Heart of*

It All: The Concept of Presence. In the Journal of Computer Mediated Communication, JCMC, 3(2).

Slater, M., Linakis, V., Usoh, M., and Kooper, R. (1996). *Immersion, Presence, and Performance in Virtual Environments: An Experiment with Tri-Dimensional Chess*. In ACM VR Software and Technology, VRST, Green, M. [Ed.], pp163-172.

French-English Literary Translation Aided by Frequency Comparisons from ARTFL and Other Corpora

Joel GOLDFIELD

Modern Languages and Literatures, Fairfield University

This presentation proposes a procedure for using frequency comparisons to help resolve challenging word choices in French-to-English literary translations. It explores a computer-assisted approach for enhancing the human translation of literary texts by two principal and complementary means: 1) through the comparison of existing translations, when available, as metatranslational, cognitive choices; 2) through the interlinguistic comparison by word frequency of cognitively admissible word choices ostensibly available to the source-language (SL) author and the chronologically distanced target-language (TL) translator. The methodology explored here does not purport to innovate in regards to machine translation but rather attempts to show how techniques in part developed by researchers in that field can assist human translators working with literature, an area where machine translation would not normally be used.

In translating "L'Illustre Magicien" (The Illustrious Magician) and "L'Histoire de Gambèr-Aly" (The Story of Gamber Aly), of Arthur de Gobineau (1816-1882), the presenter compares word frequencies in both languages—French and English—to help determine the most suitable choice where several reasonable ones exist. This procedure consists of what he calls interlingual and intralingual frequency comparisons which expand on the concept of componential analysis (CA) proposed by Newmark (Approaches to Translation, 1981) which the latter proposed as an improvement on the matrix method, also addressed by Hervey and Higgins (Thinking French Translation, 2nd ed., 2002).

In one example of his CA, Newmark develops an "...'open' series of words...and the use and choice of such words is determined as often by appropriate collocation

as by intrinsic meaning (i.e. componential analysis): this particularly applies to generic terms or head-words such as 'big' and 'large', which are difficult to analyze" (29). Newmark also notes that the word-series he chooses (bawdy, ribald, smutty, lewd, coarse, etc.) creates a problem in that it is particularly "...closely linked to any SL and TL culture in period of time and social class...."

While the approach proposed here does not distinguish social class, the interlingual frequency comparisons can usually rely on corpora and subsets established from literature written in the same time period as the works being analyzed and as early translations. In this case, several different corpora are used. For French: ARTFL (American and French Research on a Treasury of the French Language) and other tools made available by Etienne Brunet and the University of Nice. For English: Bartleby; the British National Corpus (BNC), the British Women Writers Project and Chadwyk-Healy's LION (Literature Online).

Besides the two translations, the overall project includes critical essays which treat the tales' major themes of love, death, and intellectual or emotional obsession. Honoré de Balzac's La recherche de l'Absolu (1834) floats tempingly in the background of "The Illustrious Magician" since the latter's author even published an essay on Balzac in 1844. While several different examples will be used for applying the frequency approach, one example of using interlinguistic frequencies occurs when translating the following French from "The Illustrious Magician", a several pages before its conclusion: "En effet, en entrant dans une des grottes, après en avoir visité deux ou trois, il aperçut son maître assis sur une pierre, et traçant avec le bout de son bâton des lignes, dont les combinaisons savantes annonçaient un travail divinatoire" (Nouvelles asiatiques, 1876, p. 150). Focusing for the purposes of this abstract on the clause containing the word "divinatoire," we can initially compare the two existing translations. Both were published in New York, the first by Appleton Press in 1878, the second by Helen Morganthau Fox with Harcourt Brace in 1926:

1) Appleton (259): "...the profound combination of which announced a work of divination."

2) Fox (268): "...the learned combinations of which showed that it was a work of divination."

3) Draft of presenter's translation: " whose learned combinations revealed a divinatory work."

One might wonder whether the word "divinatoire" for a French writer in the 1870's is as recherché as the word "divinatory" in English. Would it be reasonable to substitute the collocation "divinatory work" for the earlier one, "work of divination"?

The BNC reveals two (2) occurrences of «divinatory» in 100,000,000 words: 1) A6C Seeing in the dark. ed. Breakwell, Ian and Hammond, Paul. London: Serpent's Tail, 1990 (32,621 words); 2) CS0 Social anthropology in perspective. Lewis, I M. Cambridge: Cambridge University Press, 1992, pp. 5-130, (35,946 words). In the British Women Writers Project (http://www.lib.ucdavis.edu/English/BWRP/) we find no occurrences for eighty (80) texts, 1789-1832. We should conclude that the word "divinatory" is a rare word in English. However, within the ARTFL database of French works for the years 1850-1874 alone, out of 11,214,324 words, there are four (4) occurrences. Out of the 9,548,198 words in the database for the period 1875-1899 there are ten (10) occurrences or a total of 14 in 20,762,522 words or about 67 occurrences per 100,000,000 words, about 30 times more than in the BNC. In the period 1875-1899, its rate of occurrence was about once per million words, the highest of the four quarters in the century (see http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/freqs/tlf.timerseries.html). We must conclude that the French use of "divinatoire" at the time when Gobineau was using it was significantly more common than is its English counterpart though both uses are nonetheless relatively rare. And the bibliography for the fourteen occurrences covering the period in which the Nouvelles asiatiques was published includes works by well-known authors such as Amiel, Bourget, Flaubert, Garnier, the Goncourts, Mallarmé and others, as well as Gobineau himself. The one occurrence of «divinatoire» from this last author's work should of course be subtracted from the total before any comparison with the same publication.

Intralinguistically for French, there are forty-two (42) occurrences of the word "divination," which is exactly three times more frequent than "divinatoire" in the period 1850-1899, out of almost 21 million words. Forty-two versus fourteen in that number is almost a third of an order of magnitude and worthy of notice. We can preserve the rarer use of divinatoire from French in the translation although the usage appears to be even rarer in English. However, we should note that there are seven (7) occurrences of "divinatory" in the 20th-century

poetry on Chadwyk Healy's LION site, so the word can be well represented in the poetic genre. And in comparisons that will be made in the presentation in both vocabulary and themes between Gobineau and Balzac, the interest of both authors in divination will be highlighted. Brunet's database shows seven occurrences of divination/ divinations in the Comédie Humaine (CH) and four of divinatoire, a typical total of the two words (11) for the 4,242,038 words in that corpus (CH) and for his time, 1825-1849: 35 occurrences in the 12,352,370 words contained in the ARTFL database. However, these words are only about one-half as common in French literature for the last quarter of the nineteenth century as in the second quarter. There are 18 occurrences in 9,548,198 words for the ARTFL database during the period 1875-1899 when the Nouvelles asiatiques were published.

Besides the few sample words above where the computer-assisted techniques have been applied by a human translator of French literature, the presentation will suggest how such frequency-based comparisons can assist in the translation of thematic groups of words representing the literary authors' symbolic universes. Often such clusters of associated words can be determined from methodically searching the secondary literature: the thematic areas where critics have focused their interest over the centuries. Furthermore, a complementary technique for using advanced search engines on the Internet to aid in solving translation problems will be illustrated or demonstrated.

As long as French and English databases remain available with tools that allow for the appropriate date-stamping, so to speak, of word usage, a methodology can be developed using frequencies and simple statistical tests such as z-scores for comparing the ranking of words across chronological gaps. Such resources offer new and useful tools to the translator both in aiding the development of metatranslations and justifying both them and final translations. Additionally, this process can facilitate greater detail and support for literary criticism that makes use of intertextual and intratextual linguistic materials.

Stylometry, Chronology and the Styles of Henry James

David L. HOOVER

Department of English, New York University

Contemporary stylistic and stylometric studies usually focus on an author with a distinctive style and often characterize that style by comparing the author's texts to those of other authors. When an author's works display diverse styles, however, the style of one text rather than the style of the author becomes the appropriate focus. Because authorship attribution techniques are founded upon the premise that some elements of authorial style are so routinized and habitual as to be outside the author's control, extreme style variation within the works of a single author seems to threaten the validity of the entire enterprise. This apparent contradiction is only apparent, however, for the tasks are quite different. Successful attribution of a diverse group of texts to their authors requires only that each author's texts be more similar to each other than they are to texts by other authors, or, perhaps more accurately, that they be less different from each other than from the other texts. The successful separation of texts or sections of texts with distinctive styles from the rest of the works of an author takes for granted a pool of authorial similarities and isolates whatever differences remain.

Recent work has shown that the same techniques that are able to attribute texts correctly to their authors even when some of the authors' styles are quite diverse do a good job of distinguishing an unusual passage within a novel from the rest of the text (Hoover, 2003). Other quite subtle questions have also been approached using authorship attribution techniques. Nearly 20 years ago, Burrows showed that Jane Austen's characters can be distinguished by the frequencies of very frequent words in their dialogue (1987). More recent studies have used authorship techniques to investigate the sub-genres and varied narrative styles within Joyce's *Ulysses* (McKenna and Antonia, 2001), the styles of Charles Brockden Brown's narrators (Stewart, 2003), a parody of Richardson's *Pamela* (Burrows, 2005), and two translations of a Polish trilogy made a hundred years apart (Rybicki, 2005). Hugh Craig has investigated

chronological changes in Ben Jonson's style (1999a, 1999b), and Burrows has discussed chronological changes in the novel genre (1992a).

I am using authorship attribution techniques to study the often-remarked differences between Henry James's early and late styles.[1] I begin by analyzing a corpus of 46 American novels of the late 19th and early 20th century (12 by James and 34 by eight other authors) to determine the extent to which multivariate authorship attribution techniques based on frequent words, such as principal components analysis, cluster analysis, Burrows's Delta, and my own Delta Primes, successfully attribute James's early and late novels to him and distinguish them from novels by eight of his contemporaries.[2] Because all of these techniques are very effective in this task, all are appropriate for further investigation of the variation within James's style, but DeltaLz produces especially accurate results, correctly attributing all 40 novels by members of the primary set in eleven analyses based on the 2000-4000 most frequent words. All of the results also reconfirm recent findings that large numbers of frequent words are more effective than the 50-100 that have been traditionally used, and that the most accurate results (for novel-sized texts) often occur with word lists of more than 1000 words (see Hoover, 2004a, 2004b). The PCA analysis in Fig. 1, based on the 1001-1990 most frequent words, clusters the novels quite well–better, in fact, than analyses that include the 1000 most frequent words.

When cluster analysis, PCA, Delta, and Delta Prime techniques are applied to nineteen novels by Henry James, they show that the early (1971-1881) and late styles (1897-1904) are very distinct indeed, and that an "intermediate" style (1886-1890) can also be distinguished. DeltaLz again produces especially accurate results, correctly identifying all early, intermediate, and late novels in 24 analyses based on the 200-4000 most frequent words. These results paint a remarkable picture of an author whose style was constantly and consistently developing, a picture that is congruent with James's reputation as a meticulous craftsman who self-consciously transformed his style over his long career. A comparison with Charles Dickens and Willa Cather shows that Dickens's early and late novels tend to separate, but do not fall into such neat groups as James's do, and that Cather's novels form consistent groupings that are not chronological. These authors seem not to have experienced the kind of progressive development seen in James. It is dangerous, then, simply to assume chronological development of authorial style.

Finally, these same techniques show that the heavily revised versions of *The American* (1877), *Daisy Miller* (1878), and *The Portrait of a Lady* (1881) that appear in the New York edition of James's novels (1907-09) are consistently and dramatically closer to the style of the later novels. Yet even his notoriously detailed and extensive revisions do not allow PCA to group the revised early novels with the late novels. Instead, the revised versions fall at the border between the early and intermediate novels in PCA graphs (see Fig. 2), and consistently join with their original versions in cluster analyses. The results obtained using Delta show that even the errors make sense. In analyses that are not completely accurate, *The Portrait of a Lady* and *Washington Square,* the latest of the early novels (both 1881) are sometimes identified as intermediate. The other errors involve the identification of *The Spoils of Pointon,* the first of the late novels (1897), as intermediate; no analyses incorrectly identify an early novel as late or a late novel as early. In the analyses that are completely correct for the 19 unambiguously early, intermediate and late novels, the New York edition versions of earlier novels are identified as follows: the revised early novels *The American* and *Daisy Miller* are universally identified as early, the revised intermediate novel *The Reverberator* is universally labeled intermediate, and the revised early *The Portrait of a Lady* is usually labeled intermediate, but sometimes early. This is an intuitively plausible result, with the latest of the early novels pulled far enough toward the late novels to appear intermediate, but, interestingly enough, DeltaLz, which produces much more accurate results overall, labels all of the novels according to their original publication dates in eighteen analyses based on the 200-2800 mfw. In the remaining six analyses, the early *Daisy Miller* is identified as late, and in one analysis (based on the 4000mfw) *The Portrait of a Lady* is identified as intermediate. Further investigation of the implications of these results is ongoing.

Authorship attribution techniques thus confirm the traditional distinction between early and late James, establish the existence of an intermediate style, and lay the groundwork for a fuller analysis of the linguistic and stylistic differences upon which they rest. Based as they are on a very large proportion of the text of the novels (the 4000 most frequent words typically comprise more than 94% of a novel), these results provide a wealth of

material for stylistic analysis. The huge numbers of words involved will, however, require new methods of selection, analysis, and presentation if they are not to prove overwhelming and incomprehensible. Meeting these challenges will advance and refine authorship attribution techniques, and, at the same time, further illuminate the linguistic bases of James's style and his process of revision.

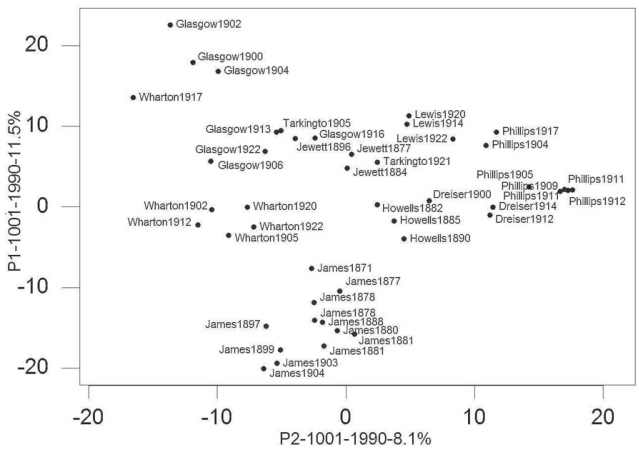


Fig. 1. 46 Novels by Henry James and 8 Other Authors

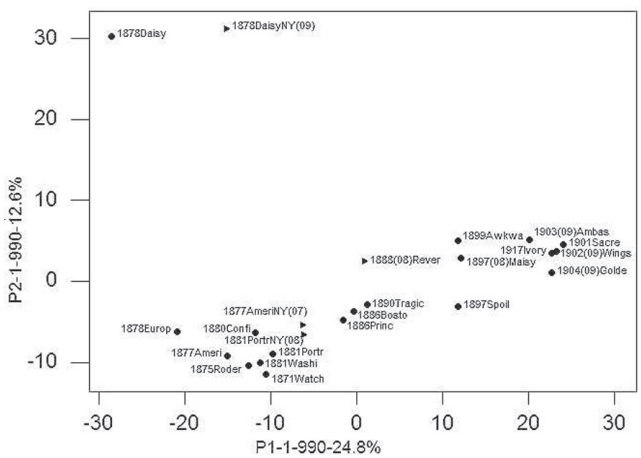


Fig. 2. Twenty-Three Novels by Henry James (revised versions marked with triangles)

References

Burrows, J. F. (1987). *Computation into Criticism. Oxford:* Clarendon Press.

Burrows, J. F. (1992a). "Computers and the study of literature." In Butler, C. S., ed. *Computers and Written Texts*. *Oxford: Blackwell,* 167-204.

Burrows, J. F. (1992b). "Not unless you ask nicely: the interpretative nexus between analysis and information," *LLC 7:* 91-109.

Burrows, J. F. (2002a). "'Delta': a measure of stylistic difference and a guide to likely authorship". LLC 17: 267-287.

Burrows, J. F. (2002b). "The Englishing of Juvenal: computational stylistics and translated texts." Style 36: 677-99.

Burrows, J. F. (2003). "Questions of authorship: attribution and beyond." *CHUM* 37: 5-32.

Burrows, J. F. (2005). "Who wrote Shamela? Verifying the authorship of a parodic text," *LLC* 20: 437-450.

Craig, H. (1999a). "Contrast and change in the idiolects of Ben Jonson characters," *CHUM* 33: 221-240.

Craig, H. (1999b). "Jonsonian chronology and the styles of a Tale of a Tub. In Butler, M. (ed.). *Re-Presenting Ben Jonson: Text, History, Performance*. Macmillan. St. Martin's, Houndmills, England, 210-32.

Hoover, D. L. (2003). "Multivariate analysis and the study of style variation." *LLC* 18: 341-60.

Hoover, D. L. (2004a). "Testing Burrows's Delta." *LLC* 19: 453-475.

Hoover, D. L. (2004b). "Delta prime?" *LLC* 19: 477-495.

McKenna, C. W. F. and A. Antonia. (2001). "The statistical analysis of style: reflections on form, meaning, and ideology in the 'Nausicaa' episode of *Ulysses," LLC* 16: 353-373.

Rybicki, J. (2005). "Burrowing into translation: character idiolects in Henryk Sienkiewicz's trilogy and its two English translations," *LLC* Advance Access published on March 24, 2005. doi:10.1093/llc/fqh051

Stewart, L. (2003). "Charles Brockden Brown: quantitative analysis and literary interpretation," *LLC* 18: 129-138.

"Quite Right, Dear and Interesting": Seeking the Sentimental in Nineteenth Century American Fiction

Tom HORTON
Kristen TAYLOR
University of Virginia

Bei YU
Xin XIANG
University of Illinois at Urbana-Champaign

This paper connects research by the nora Project (http://noraproject.org), a study on text mining and humanities databases that includes four sites and scholars from many areas, with current critical interests in nineteenth century American sentimental literature.

The term sentimental has been claimed and disparagingly applied (sometimes simultaneously) to popular fiction in this time period since its publication; academic study of sentimental fiction has enjoyed widespread acceptance in literature departments only in the past few decades. Academic disagreement persists about what constitutes sentimentality, how to include sentimental texts on nineteenth century American syllabi, which sentimental texts to include, and how to examine sentimental texts in serious criticism. Most of the well-known and widely-taught novels of the time period exist in XML format in the University of Virginia's Etext Center, one of the libraries in partnership with the nora Project; the original XML data for three texts discussed below was taken from this source.

The term sentimental novel is first applied to eighteenth century texts such as Henry Mackenzie's Man of Feeling, Samuel Richardson's Pamela, and Lawrence Sterne's Sentimental Journey and Tristam Shandy. Usually included in courses on the theory of the novel or eighteenth century literature, these works illustrate the solidification of the novel form. Sentimental novels emphasize, like Mackenzie's title, (men and women of) feeling. Feeling is valued over reason and sentimental is used with the term sensibility (recall Jane Austen's title Sense and Sensibility.) Although definitions of sentimentality range widely, and are complicated by the derogatory deployment of the term by contemporary and current critics, the group of texts loosely joined as being in the mid-nineteenth century sentimental period is a crucial link for humanities scholars that work on the novel and British and American texts in the nineteenth century connecting Victorian texts with their predecessors.

Sentimental texts are a particularly good place to look at how a group of texts may exhibit certain recognizable features; sentimental fiction uses conventional plot development, stock characters, and didactic authorial interventions. The emphasis is the exposure of how a text works to induce specific responses in the reader (these include psychophysiological responses such as crying and a resolve to do cultural work for nineteenth century causes such as temperance, anti-slavery, female education, and labor rights); readers do not expect to be surprised. Instead, readers encounter certain keywords in a certain order for a sentimental text to build the expected response. Although the term sentimental is used to represent an area of study and title literature courses, there is no set canon of sentimental texts because scholars do not agree on what constitutes textual sentimentality. Using text-mining on texts generally considered to exhibit sentimental features may help visualize levels of textual sentimentality in these texts and ultimately measure sentimentality in any text.

Two groups of humanist scholars scored three chapters in Harriet Beecher Stowe's text, Uncle Tom's Cabin, the most well-known and critically-acknowledged text in the group often considered sentimental. Although chapters in this text may be quite long and contain varying levels of sentimentality, the chapter as a unit was preferred as the original division structure of the text and the fact that humanist scholars expect this division and assign class reading and research by chapter units. UTC was later adapted into theatrical productions, and the idea of scenes (within chapters) may be a fruitful place to begin studying the sentimental fluctuations with a chapter unit in later phases of this project.

For the initial rubric, though, chapters were scored on a scale of 1 to 10: low is 1-3, medium 4-6, high 7-10. 10 is considered a "perfect"ly sentimental score, and as such, is only to be used when the peak of sentimental

conventions is exhibited: a character nears death and expires in a room usually full of flowers and mourners who often "swoon." The training set for this experiment includes two other texts that were scored on the same sentimentality scale, Susanna Rowson's 1794 novel Charlotte: A Tale of Truth and Harriet Jacobs's Incidents in the Life of a Slave Girl.

Since these texts were considered sentimental, most chapters were scored in the medium or high range, so the categories were changed to "highly sentimental" and "not highly sentimental." With D2K, the Naive Bayes method was used to extract features from these texts, which we might call markers of sentimentality. Looking at the top 100 of these features, some interesting patterns have emerged, including the privileging of proper names of minor characters in chapters that ranked as highly sentimental. Also interesting are blocks of markers that appear equally prevalent, or equally sentimental, we might say: numbers 70-74 are "wet," lamentations," "cheerfulness," "slave-trade," and "author." The line of critical argument that goes that the sentimental works focus on motherhood is borne out by "mother" at number 16 and "father" not in the top 100.

As we move into the next three phases of the project, we will include stemming as an area of interest in classifying the results. Phase two will use two more novels by the same authors as those in the training set; phrase three may include ephemera, broadsides, and other materials collected in the EAF collection at the UVa Etext Center. Phase four will run the software on texts considered non-sentimental in the nineteenth century and other phases might include twentieth and twenty-first century novels that are or are not considered sentimental. We hope to discover markers that can identify elements of the sentimental in any text.

Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters

Sobhan HOTA
Shlomo ARGAMON
Department of Computer Science, Illinois Institute of Technology

Moshe KOPPEL
Iris ZIGDON
Department of Computer Science, Bar-Ilan University

1. Introduction

A recent development in the study of language and gender is the use of automated text classification methods to examine how men and women might use language differently. Such work on classifying texts by gender has achieved accuracy rates of 70-80% for texts of different types (e-mail, novels, non-fiction articles), indicating that noticeable differences exist (de Vel et al. 2002; Argamon et al. 2003).

More to the point, though, is the fact that the distinguishing language features that emerge from these studies are consistent, both with each other, as well as with other studies on language and gender. De Vel et al. (2002) point out that men prefer 'report talk', which signifies more independence and proactivity, while women tend to prefer 'rapport talk' which means agreeing, understanding and supporting attitudes in situations. Work on more formal texts from the British National Corpus (Argamon et al. 03) similarly shows that the male indicators are mainly noun specifiers (determiners, numbers, adjectives, prepositions, and post-modifiers) indicating an 'informational style', while female indicators are a variety of features indicating an 'involved' style (explicit negation, first- and second-person pronouns, present tense verbs, and the prepositions "for" and "with").

Our goal is to extend this research for analyzing the relation of language use and gender for literary characters. To the best of our knowledge, there has been little work on understanding how novelists and playwrights

portray (if they do) differential language use by literary characters of different genders. To apply automated analysis techniques, we need a clean separation of the speech of different characters in a literary work. In novels, such speech is integrated into the text and difficult to extract automatically. To carry out such research, we prefer source texts which give easy access to such structural information; hence, we focus on analyzing characters in plays. The natural choice for a starting point is the corpus of Shakespeare's plays.

We thus ask the following questions. Can the gender of Shakespeare's characters be determined from their word usage? If we are able to find such word use, can we glean any insight into how Shakespeare portrays maleness and femaleness? Are the differences (if any) between male and female language in Shakespeare's characters similar to those found in modern texts by male and female authors? Can we expect the same kind of analysis in understanding Shakespeare's characters' gender, to the ones we discussed above? Keep in mind that here we examine text written by one individual (Shakespeare) meant to express words of different individuals with differing genders, as opposed to texts actually by individuals of different genders.

To address these questions, we applied text classification methods using machine learning. High classification accuracy, if achieved, will show that Shakespeare used different language for his male and for his female characters. If this is the case, then examination of the most important discriminating features should give some insight into such differences and to relate them to previous work on male/female language. The general approach of our work is to achieve a reasonable accuracy using different lexical features of the characters' speeches as input to machine learning and then to study those features that are most important for discriminating character gender.

2. Corpus Construction

We constructed a corpus of characters' speeches from 34 of Shakespearean plays, starting with the texts from the Moby Shakespeare[1]. The reason behind choosing this edition is that it is readily available on the web and has a convenient hierarchical form of acts and scenes for every play, while we do not expect editorial influence to unduly affect our differential analysis. The files collected from this web resource were converted into text files from hypertext media and then we cleaned the text files by removing stage directions. The gender of each character was entered manually. A text file for each character in each play was constructed by concatenating all of that character's speeches in the play. We only considered characters with 200 or more words. From that collection, all female characters were chosen. Then we took the same number of male characters as female characters from a play, restricted to those not longer than the longest female character from that particular play. In this way, we balanced the corpus for gender, giving a total of 83 female characters and 83 male characters, with equal numbers of males and females from each play. This corpus is termed the 'First Corpus'. We also built a second corpus based on the reviewer's comments, in which we equalized the number of words in male and female characters by taking every female character with more than 200 words and an equal number of the longest male characters from each play. The longest male and female characters were then matched for length by keeping a prefix of the longer part (male or female) of the same length (in words) as the shorter part. This procedure ensured that the numbers of words per play for both genders are exactly the same. This corpus is termed the 'Second Corpus'. We also split each corpus (somewhat arbitrarily) into 'early' and 'late' characters. We used the term early to those plays which were written in 16th century and late to those in 17th century. This chronology in plays as captured from Wikipedia[1]. The numbers of characters from each play for 'First Corpus' and 'Second Corpus' are shown in Table 1.

3. Feature Extraction

We processed the text using the ATMan system, a text processing system in Java that we have developed[3]. The text is tokenized and the system produces a sequence of tokens, each corresponds to a word in the input text file. We use two sets of words as features. A stylistic feature set (FW) is a list of more-or-less content-independent words comprising mainly function words, numbers, prepositions, and some common contractions (e.g., "you'll", "he'll"). A content-based feature set comprises all words that occur more than ten times in a corpus, termed Bag of Words (BoW).

We calculate the frequencies of these FWs and BoWs and turn them into numeric values by computing their relative frequencies, computed as follows. We first count the number of times two different features occurring together; then we divide this number to the count of the

feature in reference. In this way we calculate the relative frequency for each feature and a collection forms a feature vector, which represents a document (i.e. a character's speech). The FW set has 645 features including contractions; the BoW set has 2129 features collected from the first corpus and 2002 BoW features collected from the second corpus. The numeric vectors collected for each document is used as an input for machine learning.

4. Text Classification

The classification learning phase of this task is carried out by Weka's (Frank & Witten 1999) implementation of Sequential Minimal Optimization (Platt 1998) (SMO) using a linear kernel and default parameters. The output of SMO is a model linearly weighting the various text features (FW or BoW). Testing was done via 10 fold cross validation. This provides an estimation of generalization accuracy by dividing the corpus into 10 different subsets. The learning is then run ten times, each time using a different subset as a test set and combining the other nine subsets for training. In this way we ensure that each character is tested on at least once with training that does not include it. Tables 3 and 4 present the results obtained by running various experiments. It is clear that BoW has performed better than the FW in both selection criteria, as expected, since it has more features on which to operate. This shows that both style and content differ between male and female characters. As expected, the FWs have proven the stylistic evidence and not the content, which are visible from the Table 4. BoW gives a high 74.09 on over all corpuses with the equalizing on number of words selection strategy. Interestingly, FW gives highest accuracy of 74.28 in Late plays with only 63 training samples. This indicates that there is a greater stylistic difference between the genders in late Shakespeare than in early Shakespeare.

5. Discussion

The feature analysis phase is carried out by taking the results obtained from Weka's implementation of SMO. SMO provides weights to the features corresponding to both class labels. After sorting the features based on their weights, we collected the top twenty features from both character genders. Tables 5-8 lists the top 20 features from male and female characters and is shown with their assigned weights given by the SMO, for FWs and BoWs respectively. Tables 9-12 list the same for the Second Corpus. These tables also show the 'Average frequency of 100 words', which finds the frequency of a particular feature divided by total gender characters, and then for easy readability this figure is scaled by 100 times. To discriminate binary class labels, SMO uses positive and negative weight values in Weka's implementation. We see from the Tables 5-10, male features are designated as negative weights and female characters are given as positive weights. In top 20 male features, this can be observed that 'Average Frequency of 100 Words' value of male is more than the corresponding value for female. This hold same in the case of the top 20 female features where female 'Average Frequency of 100 Words' value is more than the male for the same feature.

Feature Analysis: BoW

We can see cardinal number usage is found in male characters. Plural and mass nouns ('swords', 'dogs', 'water') are used more in males than females. On the other hand, there is strong evidence for singular noun ('woman', 'mother', 'heart') usage in females. The use of 'prithee' as an interjection is found in female character. This may represent a politeness aspect in their attitude. The past participle form is generally found in females ('gone', 'named', 'known'). Present tense verb forms ('pour', 'praise', 'pray', 'love', 'dispatch', 'despair') are used in female characters. In the case of male characters, Shakespeare used these verb forms ('avoid', 'fight', 'wrought'). Male characters seem to be aggressive while female characters seem to be projected as supporters of relationships.

Feature Analysis: FW

We observed that Shakespeare's female characters used more adverbs and adjectives, as well as auxiliary verbs and pronouns. On the other hand, cardinal numbers, determiners, and some prepositions are generally indicative of male characters. These observations are in line with previous work (Argamon et al. 2003) on discriminating author gender in modern texts, supporting the idea that the playwright projects characters' gender in a manner consistent with authorial gender projection. We did observe some contrasting results in the FW features from the second corpus. Number (i.e. twice) is found in female characters. Certain prepositions are used for females, while negation only appears distinctive for early females. Determiner 'the' which is a strong male character indicator in first corpus is found only in early part of second corpus. Some negation ('cannot') is found in late males as well. Clearly, more and deeper analysis is needed.

6. Conclusion

This is the first work, to our knowledge, in analyzing literary character's gender from plays. It seems clear that male and female language in Shakespeare's characters is similar to that found in modern texts by male and female authors (Argamon et.al 2003), but more work is needed in understanding character gender. We have also observed possible differences between early and late Shakespeare in gender character classification. In particular, the later Shakespeare plays appear to show a greater stylistic discrimination between male and female characters than the earlier plays. We are particularly interested in collaborating with literary scholars on this research to explore these issues further.

Play Name	Gender Count
All's Well That Ends Well	8
Antony and Cleopatra	4
As You Like It	6
Cymbeline	4
King Lear	4
Loves Labours Lost	8
Measure for Measure	4
Midsummer Nights Dream	6
Much Ado About Nothing	8
Othello The Moore of Venice	6
Pericles Prince of Tyre	8
Romeo and Juliet	6
The Comedy of Errors	8
The First part of King Henry The Fourth	2
The First part of King Henry The Sixth	4
The Life and Death of Julies Caesar	2
The Life and Death of Richard The Second	4
The Life and Death of Richard The Third	8
The Life of King Henry The Eighth	4
The Life of King Henry The Fifth	4
The Merchant of Venice	6
The Merry Wives of Windsor	6
The Second part of King Henry The Fourth	2
The Second part of King Henry The Sixth	2
The Taming of the Shrew	4
The Tempest	2
The Third part of King Henry The Sixth	6
The Tragedy of Coriolanus	4
The Tragedy of Hamlet	2
Titus Andronicus	4
Troilus and Cressida	2
Twelfth Night	6
Two Gentlemen of Verona	6
Winter's Tale	6

Table 1: Corpus Composition

	Male	Female
All	83	83
Early	48	48
Late	35	35

Table 2: Overall Corpus Statistics

Feature Set	Accuracy
All	
Function Words	**66.26**
Bag-of-Words	**73.49**
Early	
Function Words	**63.54**
Bag-of-Words	**62.50**
Late	
Function Words	**62.85**
Bag-of-Words	**60.00**

Table 3: Accuracy is expressed in percentage for First Corpus Selection

Feature Set	Accuracy
All	
Function Words	65.66
Bag-of-Words	74.09
Early	
Function Words	56.25
Bag-of-Words	58.33
Late	
Function Words	74.28
Bag-of-Words	64.28

Table 4: Accuracy is expressed in percentage for Second Corpus Selection

Features from Various Experiments Using First Corpus

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
this	-0.7154	1195.18	845.78	look	0.7082	174.69	159.03
follows	-0.658	12.04	3.61	such	0.6609	308.43	240.96
allow	-0.6545	4.81	1.20	thorou -ghly	0.6248	3.61	0.0
in	-0.6309	2045.78	1679.51	comes	0.6134	102.40	78.31
well	-0.6051	213.25	172.28	gone	0.6095	31.32	19.27
three	-0.5525	85.54	18.07	he's	0.5935	67.46	54.21

there	-0.5513	296.38	231.32	never	0.5841	228.91	163.85
allows	-0.5279	1.20	1.20	only	0.5805	56.62	44.57
the	-0.5184	5408.43	3906.02	am	0.5404	474.69	395.18
toward	-0.5184	14.45	4.81	he	0.5353	1198.79	1103.61
one	-0.5084	333.73	263.85	there -fore	0.5212	104.81	72.28
immed-iate	-0.4997	6.02	0.0	might	0.4557	102.40	73.49
here's	-0.4789	39.75	24.09	you	0.4375	2283.13	2116.86
appear	-0.4752	34.93	12.04	further-more	0.4349	1.20	0.0
himself	-0.4713	51.80	27.71	outside	0.4342	2.40	1.20
we'll	-0.4573	43.37	37.34	take	0.4298	237.34	228.91
another	-0.4517	66.26	40.96	brief	0.4242	8.43	4.81
five	-0.4204	39.75	8.43	you'll	0.4201	46.98	31.32
thus	-0.4201	103.61	77.10	wish	0.418	44.57	25.30
thank	-0.4122	79.51	59.03	consid-ering	0.3975	2.40	0.0

Table 5: Statistics of Top 20 FW Features from Gender Char

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
three	-0.1774	85.54	18.07	alas	0.22	12.04	1.20
lying	-0.1602	4.81	0.0	gone	0.1634	31.32	19.27
friendship	-0.1583	7.22	1.20	brow	0.1581	9.63	4.81
shortly	-0.1474	6.02	1.20	love	0.1525	343.37	209.63
bare	-0.1449	12.04	2.40	o	0.1483	279.51	172.28
wrought	-0.1438	10.84	1.20	prithee	0.1437	12.04	2.40
avoid	-0.1369	9.63	1.20	he	0.1337	1198.79	1103.61
this	-0.1333	1195.18	845.78	pray	0.1289	189.15	139.75
answer	-0.1302	63.85	33.73	dispatch	0.1289	7.22	2.40
very	-0.1255	222.89	131.32	sick	0.1275	24.09	3.61
purse	-0.1255	13.25	2.40	such	0.1273	308.43	240.96
served	-0.1217	8.43	3.61	woman	0.1268	54.21	24.09
savage	-0.1211	8.43	3.61	am	0.1257	474.69	395.18
thrice	-0.1182	12.04	7.22	glass	0.1253	7.22	3.61
whom	-0.1167	87.95	37.34	mother	0.1193	46.98	16.86
fresh	-0.1165	19.27	8.43	warrant	0.1164	45.78	24.09
her	-0.1162	736.14	536.14	colour	0.1147	12.04	6.02
fears	-0.114	8.43	1.20	me	0.1144	1046.98	1032.53
dogs	-0.1126	7.22	1.20	poor	0.1141	160.24	97.59
hundred	-0.1117	27.71	8.43	woo	0.1134	16.86	7.22

Table 6: Statistics of Top 20 BoW Features from Gender Character

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
three	-0.4626	85.54	18.07	theref-ore	0.5684	104.81	72.28
in	-0.4327	2045.78	1679.51	only	0.4555	56.62	44.57
this	-0.4065	1195.18	845.78	just	0.4092	22.89	18.07
the	-0.3971	5408.43	3906.02	thorou-ghly	0.4057	3.61	0.0
five	-0.3785	39.75	8.43	brief	0.4043	8.43	4.81
there	-0.3772	296.38	231.32	he's	0.3834	67.46	54.21
certain	-0.3691	36.14	7.22	gone	0.3657	31.32	19.27
together	-0.3684	20.48	10.84	below	0.3251	6.02	2.40

allow	-0.3609	4.81	1.20	wish	0.3177	44.57	25.30
on't	-0.359	3.61	1.20	never	0.3012	228.91	163.85
here's	-0.3537	39.75	24.09	outside	0.2931	2.40	1.20
we	-0.3247	595.18	381.92	in't	0.2931	6.02	4.81
whence	-0.3195	21.68	4.81	known	0.2784	31.32	28.91
appear	-0.3095	34.93	12.04	still	0.2719	84.81	63.85
necess-ary	-0.2927	2.40	0.0	value	0.2683	6.02	4.81
seem	-0.2716	26.50	25.30	else	0.2622	63.85	50.60
beyond	-0.2634	10.84	4.81	keep	0.2586	97.59	75.90
indeed	-0.2607	33.73	19.27	help	0.2575	40.96	39.75
wonder	-0.25	15.66	12.04	along	0.2571	22.89	14.45
upon't	-0.2497	1.20	0.0	me	0.2566	1046.98	1032.53

Table 7: Statistics of Top 20 FW Features from Early Gender Character

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
another	-0.298	66.26	40.96	such	0.4355	308.43	240.96
follows	-0.2849	12.04	3.61	gone	0.3838	31.32	19.27
of	-0.2801	3115.66	2381.92	am	0.3198	474.69	395.18
immediate	-0.2704	6.02	0.0	you	0.2784	2283.13	2116.86
toward	-0.27	14.45	4.81	on's	0.2774	4.81	2.40
three	-0.2624	85.54	18.07	you're	0.2684	21.68	8.43
cannot	-0.2459	157.83	127.71	might	0.2522	102.40	73.49
doing	-0.2422	16.86	4.81	hence	0.2518	21.68	18.07
consider	-0.2367	6.02	2.40	apart	0.2349	3.61	2.40
where	-0.235	221.68	186.74	among	0.2305	31.32	24.09
this	-0.2314	1195.18	845.78	use	0.2254	51.80	46.98
example	-0.2249	7.22	1.20	sure	0.221	46.98	36.14
very	-0.2241	222.89	131.32	seem	0.1876	31.32	26.50
whom	-0.22	87.95	37.34	he	0.1848	1198.79	1103.61
already	-0.2135	10.84	4.81	comes	0.1806	102.40	78.31
every	-0.2124	110.84	96.38	where't	0.179	1.20	0.0
own	-0.2074	143.37	109.63	almost	0.1729	36.14	24.09
be	-0.205	1265.06	1261.44	lately	0.1713	4.81	3.61
we	-0.2018	595.18	381.92	near	0.1706	32.53	16.86
rather	-0.1974	87.95	65.06	little	0.1662	86.74	77.10

Table 8: Statistics of Top 20 FW Features from Late Gender Character

Features from Various Experiments Using Second Corpus

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
in	-0.85	1262.65	1037.95	he	0.78	1623.49	1454.81
three	-0.71	47.59	13.85	whoe'er	0.72	1.80	0.0
itself	-0.70	29.51	13.85	him	0.64	512.65	360.24

toward	-0.68	15.66	7.83	below	0.58	4.81	2.40
follows	-0.67	9.63	1.80	wish	0.57	34.93	26.50
thus	-0.63	69.87	54.21	did	0.51	158.43	103.01
allow	-0.61	8.43	3.01	he's	0.48	33.13	19.27
will	-0.57	501.20	440.36	thorough	0.47	3.01	0.60
necess-ary	-0.51	2.40	0.60	thou'dst	0.47	0.60	0.0
being	-0.49	75.90	44.57	every	0.47	46.98	39.75
beyond	-0.49	6.62	3.01	whet'st	0.47	0.60	0.0
gives	-0.47	15.06	9.03	outside	0.45	1.20	0.0
these	-0.47	109.63	87.95	gone	0.45	59.03	35.54
another	-0.46	37.34	27.71	else	0.44	40.96	31.92
may	-0.45	156.02	135.54	me	0.43	1157.83	1045.18
whence	-0.45	9.63	2.40	to's	0.43	1.80	0.60
whom	-0.45	39.15	21.68	down	0.43	51.20	39.15
allows	-0.44	0.60	0.0	help	0.43	34.93	24.09
of	-0.44	1418.67	1225.30	twice	0.42	8.43	3.61
after	-0.43	42.16	25.90	does	0.42	39.15	25.30

Table 9: Statistics of Top 20 FW Features from All Gender Character

gives	-0.45	15.06	9.03	help	0.42	34.93	24.09
after	-0.44	42.16	25.90	need	0.41	37.34	29.51
but	-0.44	519.27	551.20	whole	0.40	12.04	7.22
thus	-0.41	69.87	54.21	he's	0.38	33.13	19.27
allows	-0.41	0.60	0.0	gone	0.36	59.03	35.54
necess-ary	-0.41	2.40	0.60	wish	0.36	34.93	26.50
beyond	-0.38	6.62	3.01	he	0.35	1623.49	1454.81
greetings	-0.35	1.80	0.60	old	0.35	40.96	46.38
here	-0.35	178.31	166.26	never	0.35	109.63	80.72
got	-0.35	9.03	6.02	keep	0.35	55.42	49.39
thence	-0.33	11.44	4.81	oh	0.32	0.60	0.0
how	-0.32	200.60	179.51	will't	0.31	1.80	0.0
her	-0.31	611.44	481.32	for't	0.31	8.43	6.02
follows	-0.31	9.63	1.80	every	0.31	46.98	39.75
accord-ing	-0.31	7.83	3.01	away	0.31	60.84	53.61
certain	-0.31	16.86	13.25	thou'dst	0.29	0.60	0.0
ta'en	-0.29	7.22	5.42	him	0.29	512.65	360.24

Table 11: Statistics of Top 20 FW Features from Early Gender Character

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
in	-0.18	1262.65	1037.95	prithee	0.25	27.10	5.42
three	-0.16	47.59	13.85	him	0.23	512.65	360.24
her	-0.15	611.44	481.32	alas	0.22	17.46	2.40
will	-0.14	501.20	440.36	he	0.22	1623.49	1454.81
answer	-0.14	47.59	28.31	heart	0.21	134.93	88.55
seest	-0.14	7.83	1.20	o	0.19	2519.87	2391.56
appetite	-0.14	5.42	1.20	mother	0.17	59.03	24.09
being	-0.14	75.90	44.57	fortune	0.17	42.77	27.71
prepare	-0.14	9.03	3.61	maiden	0.16	14.45	3.01
to	-0.13	1985.54	1822.28	warrant	0.15	30.72	8.43
thrive	-0.13	9.63	1.20	master's	0.15	9.63	1.20
itself	-0.13	29.51	13.85	dispatch	0.15	9.63	3.61
hopes	-0.13	7.83	3.01	wish	0.15	34.93	26.50
months	-0.13	8.43	1.20	easily	0.14	4.21	1.80
another	-0.13	37.34	27.71	did	0.14	158.43	103.01
fellow	-0.13	35.54	20.48	named	0.13	6.02	1.20
fellows	-0.13	7.22	2.40	lord	0.13	298.19	137.95
steel	-0.12	7.22	1.80	does	0.12	39.15	25.30
ink	-0.12	6.62	1.80	pray	0.12	125.30	72.28
greatn-ess	-0.12	10.24	2.40	same	0.12	22.28	10.84

Table 10: Statistics of Top 20 BoW Features from All Gender Character

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
itself	-0.33	29.51	13.85	he	0.40	1623.49	1454.81
in	-0.30	1262.65	1037.95	who's	0.32	7.83	3.01
toward	-0.25	15.66	7.83	such	0.30	150.0	115.66
the	-0.23	3341.56	2884.93	does	0.28	39.15	25.30
three	-0.22	47.59	13.85	best	0.26	50.0	37.95
cannot	-0.21	75.90	62.65	tell	0.24	118.07	96.98
and	-0.21	1917.46	1801.80	might	0.23	60.24	43.37
of	-0.21	1418.67	1225.30	him	0.23	512.65	360.24
follows	-0.20	9.63	1.80	you	0.23	2054.81	1733.73
own	-0.20	76.50	60.24	your	0.22	665.66	585.54
ones	-0.20	7.22	3.61	last	0.20	29.51	20.48
followed	-0.2	1.80	1.20	goes	0.20	17.46	7.83
may	-0.19	156.02	135.54	among	0.20	10.84	7.83
without	-0.19	28.91	19.87	little	0.20	49.39	42.16
someb-ody	-0.19	0.60	0.0	on's	0.19	2.40	0.60
towards	-0.19	8.43	6.62	lately	0.19	2.40	1.20
this	-0.18	530.72	447.59	almost	0.19	15.66	9.63
beyond	-0.18	6.62	3.01	twice	0.17	8.43	3.61
gives	-0.18	15.06	9.03	yes	0.17	7.83	3.01
another	-0.17	37.34	27.71	howe'er	0.16	2.40	0.0

Table 12: Statistics of Top 20 FW Features from Late Gender Character

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
will	-0.58	501.20	440.36	only	0.51	28.31	24.69
in	-0.58	1262.65	1037.95	below	0.47	4.81	2.40
three	-0.51	47.59	13.85	there-fore	0.42	53.01	49.39

References

Koppel M., Argamon S., Shimoni A. (2004). Automatically

Categorizing Written Texts by Author Gender : *Literary and Linguistic Computing* 17(4).

Argamon S., **Koppel M.**, **Fine J.**, **Shimoni A**. (2003). Gender, Genre and Writing Style in Formal Written Texts : *Text* 23(3), pp. 321–346

Argamon S., Whitelaw C., Chase P., Hota S., Dhawle S., Garg N., Levitan S. (2005) *Stylistic Text Classification using Functional Lexical Features, Journal of the Association for Information Sciences and Technology*, to appear.

Corney M., Vel O., Anderson A., Mohay G. (2002). *Gender Preferential Text Mining of E-mail Discourse :* In *Proceedings of 18th Annual Computer Security Applications Conference ACSAC*

Corney M., Vel O., Anderson A. (2001). Mining E-mail Content for Author Identification Forensics *: ACM SIGMOD Record* Volume 30, Issue 4 December

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with many relevant features. *ECML-98, Tenth European Conference on Machine Learning*.

Platt, J. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Microsoft Research Technical Report MSR-TR-98-14,

Mitchell, T. (1997) *Machine Learning*. (McGraw-Hill)

Witten I., Frank E. (1999). *Weka3: Data Mining Software in Java* http://www.cs.waikato.ac.nz/ml/weka/**Tables**

Criticism Mining: Text Mining Experiments on Book, Movie and Music Reviews

Xiao HU
J. Stephen DOWNIE
M. Cameron JONES

University of Illinois at Urbana-Champaign

1. INTRODUCTION

There are many networked resources which now provide critical consumer-generated reviews of humanities materials, such as online stores, review websites, and various forums including both public and private blogs, mailing lists and wikis. Many of these reviews are quite detailed, covering not only the reviewers' personal opinions but also important background and contextual information about the works under discussion. Humanities scholars should be given the ability to easily gather up and then analytically examine these reviews to determine, for example, how users are impacted and influenced by humanities materials. Because the ever-growing volume of consumer-generated review text precludes simple manual selection, the time has come to develop robust automated techniques that assist humanities scholars in the location, organization and then the analysis of critical review content. To this end, the authors have conducted a series of very promising large-scale experiments that bring to bear powerful text mining techniques to the problem of "criticism analysis". In particular, our experimental results concerning the application of the Naïve Bayes text mining technique to the "criticism analysis" domain indicate that "criticism mining" is not only feasible but also worthy of further exploration and refinement. In short, our results suggest that the formal development of a "criticism mining" paradigm would provide humanities scholars with a sophisticated analytic toolkit that will open rewarding new avenues of investigation and insight.

2. EXPERIMENTAL SETUP

Our principal experimental goal was to build and then evaluate a prototype criticism mining system that could automatically predict the:

1) genre of the work being reviewed (Experimental Set 1 (ES1)).

2) quality rating assigned to the reviewed item (ES2).

3) difference between book reviews and movie reviews, especially for items in the same genre(ES3).

4) difference between fiction and non-fiction book reviews (ES4).

In this work, we focused on the movie, book and music reviews published on www.epinions.com, a website devoted to consumer-generated reviews. Each review in epinions.com is associated with both a genre label and a numerical quality rating expressed as a number of stars (from 1 to 5) with higher ratings indicating more positive opinions. The genre labels and the rating information provided the ground truth for the experiments. 1800 book reviews, 1650 movie reviews and 1800 music reviews were selected and downloaded from the most popular genres represented on epinions.com. As in our earlier work (Hu et al 2005), the distribution of reviews across genres and ratings was made as evenly as possible to eliminate analytic bias. Each review contains a title, the reviewer's star rating of the item, a summary, and the full review content. To make our criticism mining approach generalizable to other sources of criticism materials, we only processed the full review text and the star rating information. Figure 1 illustrates the movie, book and music genre taxonomies used in our experiments.

Books		Movies	Music
Fiction	Non-fiction		
Action & Thrillers [1]		Action/Adventure [1]	Blues
Juvenile Fiction [2]		Children [2]	Classical
	Humor [3]	Comedies [3]	Country
Horror [4]		Horror/Suspense [4]	Electronic
	Music & Performing Arts 5	Musical & Performing Arts 5	Gospel
Science Fiction & Fantasy [6]		Science-Fiction/ Fantasy 6	Hardcore / Punk
	Biography & Autobiography	Documentary	Heavy Metal
Mystery & Crime		Dramas	International
		Education/ General Interest	Jazz Instrument
Romance		Japanimation (Anime)	Pop Vocal
		War	R&B
			Rock & Pop

Figure 1: Book, movie and music genres from epinions.com used in experiments; Genres with the same superscripts are overlapping ones used in "Books vs. Movie Reviews" experiments (ES3)

The same data preprocessing and modeling techniques were applied to all experiments. HTML tags were removed, and the documents were tokenized. Stop words and punctuation marks were not stripped as previous studies suggest these provide useful stylistic information (Argamon and Levitan 2005, Stamatatos 2000). Tokens were stemmed to unify different forms of the same word (e.g., plurals). Documents were represented as vectors where each attribute value was the frequency of occurrence of a distinct term. The model selected was generated by a Naïve Bayesian text classifier which has been widely used in text mining due to its robustness and computational efficiency (Sebastiani 2002). The experiments were implemented in the Text-to-Knowledge (T2K) framework which facilitates the fast prototyping of the text mining techniques (Downie et al 2005).

3. GENRE CLASSIFICATION TESTS (ES1)

Figure 2a provides an overview of the genre classification tests. The confusion matrices (Figure 2b, 2c and 2d) illustrate which genres are more distinguishable from the others and which genres are more prone to misclassification. Bolded values represent the successful classification rate for each medium (Figure 2a) or genre (Figure 2b, 2c and 2d).

	Book	Movie	Music
Number of genres	9	11	12
Reviews in each genre	200	150	150
Term list size	41,060 terms	47,015 terms	47,864 terms
Mean of review length	1,095 words	1,514 words	1,547 words
Std Dev of review length	446 words	672 words	784 words
Mean of precision	**72.18%**	**67.70%**	**78.89%**
Std Dev of precision	1.89%	3.51%	4.11%

(a) Overview Statistics of Genre Classification Experiments

T P	Action	Bio.	Horror	Humor	Juvenile	Music	Mystery	Romance	Science
Action	**0.61**	0.01	0.06	0.01	0.02	0.03	0.20	0.05	0.02
Bio.	0.04	**0.70**	0.01	0.05	0.03	0.13	0.01	0.03	0
Horror	0.09	0	**0.66**	0	0.05	0	0.12	0.02	0.06
Humor	0.01	0.10	0	**0.74**	0.03	0.08	0.01	0.01	0.03
Juvenile	0.01	0.01	0	0.07	**0.86**	0.02	0	0.02	0
Music	0	0.09	0	0	0.01	**0.89**	0	0	0.01
Mystery	0.20	0	0.01	0	0.01	0	**0.70**	0.05	0.04
Romance	0.06	0.01	0.01	0	0.04	0	0.08	**0.78**	0.03
Science	0.03	0	0.02	0.01	0.11	0.03	0.01	0.13	**0.66**

(b) Book Review Genre Classification Confusion Matrix

T P	Action	Anime	Children	Comedy	Docu.	Drama	Edu.	Horror	Music	Science	War
Action	**0,77**	0	0	0.01	0	0.01	0.02	0	0	0.10	0.09
Anime	0	**0.89**	0.03	0.03	0	0	0	0	0	0.05	0
Children	0.02	0.01	**0.95**	0	0.01	0.01	0.01	0	0	0	0
Comedy	0.09	0.01	0.06	**0.52**	0.03	0.17	0.06	0.01	0.03	0.01	0.02
Docu.	0.02	0	0	0.04	**0.63**	0.01	0.19	0	0.09	0	0.02
Drama	0.16	0	0	0.12	0.10	**0.45**	0.05	0.03	0.03	0.01	0.04
Edu.	0	0	0.02	0.02	0.31	0.03	**0.57**	0	0	0.01	0.03
Horror	0.15	0.02	0.02	0.02	0.03	0.02	0.05	**0.69**	0	0.10	0.02
Music	0	0	0	0.01	0.18	0	0	0	**0.81**	0	0
Science	0.04	0.01	0.02	0	0.06	0.01	0.02	0.03	0	**0.76**	0.05
War	0.11	0	0.01	0.01	0.08	0.08	0.05	0.03	0.02	0.02	**0.59**

(c) Movie Review Genre Classification Confusion

T P	Blues	Classical	Country	Electr.	Gospel	Punk	Metal	Int'l	Jazz	Pop Vo.	R&B	Rock
Blues	**0.61**	0	0.10	0	0	0	0	0	0	0	0	0.29
Classical	0	**0.94**	0	0.03	0	0	0	0	0	0	0	0.03
Country	0	0	**0.92**	0	0.03	0	0	0	0	0	0	0.06
Electr.	0	0	0	**0.92**	0	0	0.06	0	0	0	0	0.03
Gospel	0	0	0.05	0	**0.80**	0	0	0	0	0	0.05	0.10
Punk	0	0	0	0.05	0	**0.71**	0.05	0	0	0	0	0.19
Metal	0	0	0	0	0	0	**0.89**	0	0	0	0	0.11
Int'l	0	0.04	0.00	0.04	0	0	0	**0.81**	0	0	0	0.04
Jazz	0	0	0	0.04	0	0	0	0	**0.89**	0.04	0	0.04
Pop Vo.	0	0	0.04	0.07	0	0	0	0.04	0.07	**0.68**	0	0.11
R&B	0	0	0	0	0	0	0	0	0	0.06	**0.88**	0.06
Rock	0.03	0	0.03	0	0	0	0.03	0	0	0.03	0	**0.89**

(d) Music Review Genre Classification Confusion Matrix

Figure 2: Genre classification data statistics, results and confusion matrices. The first rows in confusion matrices represent prediction (P); the first columns represent ground truth (T). 5- fold random cross-validation on book and movie reviews, 3- fold random cross-validation on music reviews

As Figure 2a shows, the overall precisions are impressively high (67.70% to 78.89%) compared to the baseline of random selection (11.11% to 8.33%). The identification of some genres is very reliable e.g., "Music & Performing Arts" book reviews (89%) and "Children" movie reviews (95%). Some understandable confusions are also apparent e.g., "Documentary" and "Education" movie reviews (31% confusion). High confusion values appear to indicate that such genres semantically overlap. Furthermore, such confusion values may also indicate pairs of genres that create similar impressions and impacts on users. For example, there might be a formal distinction between the "Documentary" and "Education" genres but the two genres appear to affect significant numbers of users in similar, interchangeable ways.

4. RATING CLASSIFICATION TESTS (ES2)

We first tested the classification of reviews according to quality rating as a five class problem (i.e., classification classes representing the individual

ratings (1, 2, 3, 4 and 5 stars)). Next we conducted two binary classification experiments: 1) negative and positive review "group" identification (i.e., 1 or 2 stars versus 4 or 5 stars); and 2) *ad extremis* identification (i.e., 1 star versus 5 stars). Figure 3 demonstrates the dataset statistics, corresponding results and confusion matrices.

Book Reviews			
Experiments	1 star … 5 stars	1, 2 stars vs. 4, 5 stars	1 star vs. 5 stars
Number of classes	5	2	2
Reviews in each class	200	400	300
Term list size	34,123 terms	28,339 terms	23,131 terms
Mean of review length	1,240 words	1,228 words	1,079 words
Std Dev of review length	549 words	557 words	612 words
Mean of precision	**36.70%**	**80.13%**	**80.67%**
Std Dev of precision	1.15%	4.01%	2.16%
Movie Reviews			
Experiments	1 star … 5 stars	1, 2 stars vs. 4, 5 stars	1 star vs. 5 stars
Number of classes	5	2	2
Reviews in each class	220	440	400
Term list size	40,235 terms	36,620 terms	31,277 terms
Mean of review length	1,640 words	1,645 words	1,409 words
Std Dev of review length	788 words	770 words	724 words
Mean of precision	**44.82%**	**82.27%**	**85.75%**
Std Dev of precision	2.27%	2.02%	1.20%
Music Reviews			
Experiments	1 star … 5 stars	1, 2 stars vs. 4, 5 stars	1 star vs. 5 stars
Number of classes	5	2	2
Reviews in each class	200	400	400
Term list size	35,600 terms	33,084 terms	32,563 terms
Mean of review length	1,875 words	2,032 words	1,842 words
Std Dev of review length	913 words	912 words	956 words
Mean of precision	**44.25%**	**81.25%**	**86.25%**
Std Dev of precision	2.63%	N/A	N/A

(a) Overview Statistics of Rating Classification Experiments

T P	1 star	2 stars	3 stars	4 stars	5 stars
1 star	**0.45**	0.21	0.15	0.09	0.10
2 stars	0.24	**0.36**	0.19	0.12	0.09
3 stars	0.11	0.17	**0.28**	0.22	0.21
4 stars	0.05	0.06	0.17	**0.41**	0.31
5 stars	0.04	0.07	0.17	0.26	**0.46**

(b) Book Review Rating Classification Confusion Matrix (5 ratings)

T P	1 star	2 stars	3 stars	4 stars	5 stars
1 star	**0.49**	0.19	0.17	0.08	0.07
2 stars	0.15	**0.45**	0.23	0.11	0.06
3 stars	0.04	0.24	**0.28**	0.27	0.17
4 stars	0.05	0.13	0.13	**0.41**	0.27
5 stars	0.07	0.03	0.16	0.20	**0.54**

(c) Movie Review Rating Classification Confusion Matrix (5 ratings)

T P	1 star	2 stars	3 stars	4 stars	5 stars
1 star	**0.61**	0.24	0.07	0.05	0.02
2 stars	0.24	**0.15**	0.36	0.15	0.09
3 stars	0.11	0.13	**0.41**	0.20	0.15
4 stars	0.03	0.06	0.10	**0.32**	0.48
5 stars	0	0	0.09	0.11	**0.80**

(d) Music Review Rating Classification Confusion Matrix (5 ratings)

Figure 3: Rating classification data statistics, results and confusion matrices. The first rows in confusion matrices represent prediction (P); the first columns represent ground truth (T). 5- fold random cross-validation on book and movie reviews, one single iteration on music reviews

The classification precision scores for the binary rating tasks are quite strong (80.13% to 86.25%), while the five class scores are substantially weaker (36.70% to 44.82%). However, upon examination of the five class confusion matrices it is apparent that the system is "reasonably" confusing adjacent categories (e.g., 1 star with 2 stars, 4 stars with 5 stars, etc.).

5. MOVIE VS. BOOK REVIEW TESTS (ES3)

We first formed a binary classification experiment with movie and book reviews of all genres. We then compared reviews in each of the six genres common to books and movies. To prevent the oversimplification of the classification task we eliminated words that can directly suggest the categories: "book", "movie", "fiction", "film", "novel", "actor", "actress", "read", "watch", "scene", etc. Eliminated terms were selected from those which occurred most frequently in either category but not both.

Genre	All Genres	Action	Horror	Humor/Comedy
Number of classes	2	2	2	2
Reviews in each class	800	400	400	400
Term list size	49,263 terms	24,552 terms	25,509 terms	26,713 terms
Mean of review length	1,608 words	933 words	1,779 words	1,091 words
Std Dev of review length	697 words	478 words	546 words	625 words
Mean of precision	**94.28%**	**95.63%**	**98.12%**	**99.13%**
Std Dev of precision	1.18%	0.99%	1.40%	1.05%

Genre	Juvenile Fiction /Children	Music & perorming Aarts	Science Fiction & Fantasy
Number of classes	2	2	2
Reviews in each class	400	400	400
Term list size	21,326 terms	23,217 terms	25,088 terms
Mean of review length	849 words	791 words	1,011 words
Std Dev of review length	333 words	531 words	544 words
Mean of precision	**97.87%**	**97.02%**	**97.25%**
Std Dev of precision	0.71%	1.49%	1.91%

Figure 4: Overview statistics of book and movie review classification experiments. All results are from 5 - fold random cross validation

The results in Figure 4 show the classifier is amazingly accurate (consistently above 94.28% precision) in distinguishing movie reviews from book reviews both in mixed genres and within single genre classes. We conducted a post-experiment examination of the reviews to ensure that the results were not simply based upon suggestive terms like those we had eliminated pre-experiment. Therefore, it can be inferred that users criticize books and movies in quite different ways. This is an important finding that prompts for future work the identification of key features contributing to such differences.

6. FICTION VS. NON-FICTION BOOK REVIEW TEST (ES4)

As in ES3, we eliminated such suggestive words as "fiction", "non", "novel", "character", "plot", and "story" after examining high-frequency terms of each category. The classification results are shown in Figure 5.

Fiction vs. Non-fiction	
Number of classes	2
Reviews in each class	600
Term list size	35,210 terms
Mean of review length	1,220 words
Std Dev of review length	493 words
Mean of precision	**94.67%**
Std Dev of precision	1.16%

(a) Overview Statistics of Fiction and Non-fiction Book Review Classification Experiment

T P	Fiction	Non-fiction
Fiction	**0.98**	0.02
Non-Fiction	0.09	**0.91**

(b) Fiction and Non-fiction Book Review Classification Confusion Matrix

Figure 5: Fiction and non-fiction book review classification data statistics, results and confusion matrix. The first row in confusion matrix represents prediction (P); the first column represents ground truth (T). Results are from 5- fold random cross validation

The precision of 94.67% not only verifies our system is good at this classification task but also indicates reviews on the two categories are significantly different. It is also noteworthy that more non-fiction book reviews (9%) were mistakenly predicted as fiction book reviews than the other way around (2%). Closer analysis on features causing such behaviors will be our future work.

7. CONCLUSIONS AND FUTURE WORK

Consumer-generated reviews of humanities materials represent a valuable research resource for humanities scholars. Our series of experiments on the automated classification of reviews verify that important information about the materials being reviewed can be found using text mining techniques. All our experiments were highly successful in terms of both classification accuracy and the logical placement of confusion in the confusion matrices. Thus, the development of "criticism mining" techniques based upon the relatively simple Naïve Bayes model has been shown to be simultaneously viable and robust. This finding promises to make the ever-growing consumer-generated review resources useful to humanities scholars.

In our future work, we plan to undertake a broadening of our understanding by exploring the application of text

mining techniques beyond the Naïve Bayes model (e.g., decision trees, neural nets, support vector machines, etc.). We will also work towards the development of a system to automatically mine arbitrary bodies of critical review text such as blogs, mailing lists, and wikis. We also hope to construct content and ethnographic analyses to help answer the "why" questions that pertain to the results.

References:

Argamon, **S**., and **Levitan**, **S**. (2005). Measuring the Usefulness of Function Words for Authorship Attribution. *Proceedings of the 17th Joined International Conference of ACH/ALLC*.

Downie, **J. S**., **Unsworth**, **J**., **Yu**, **B**., **Tcheng**, **D**., **Rockwell**, **G**., and **Ramsay**, **S. J**. (2005). A Revolutionary Approach to Humanities Computing?: Tools Development and the D2K Data-Mining Framework. *Proceedings of the 17th Joined International Conference of ACH/ALLC*.

Hu, **X**., **Downie**, **J. S**., **West**, K., and **Ehmann**, **A**. (2005). Mining Music Reviews: Promising Preliminary Results. *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR)*.

Sebastiani, **F**. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34, 1.

Stamatatos, E., Fakotakis, N., and **Kokkinakis, G.** (2000). Text Genre Detection Using Common Word Frequencies. *Proceedings of 18th International Conference on Computational Linguistics*.

Markup Languages for Complex Documents – an Interim Project Report

Claus HUITFELDT
Department of Philosophy, University of Bergen, Norway

Michael SPERBERG-MCQUEEN
World Wide Web Consortium, MIT Computer Science and Artificial Intelligence Laboratory (CSAIL)

David DUBIN
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Lars G. JOHNSEN
Department of Linguistics and Comparative Literature, University of Bergen, Norway

Background

Before the advent of standards for generic markup, the lack of publicly documented and generally accepted standards made exchange and reuse of electronic documents and document processing software difficult and expensive.

SGML (Standard Generalized Markup Language) became an international standard in 1986. But it was only in 1993, with the introduction of the World Wide Web and its SGML-inspired markup language HTML (Hypertext Markup Language), that generic markup started to gain widespread acceptance in networked publishing and communication.

In 1998, the World Wide Web Consortium (W3C) released XML (Extensible Markup Language). XML is a simplified subset of SGML, aimed at retaining HTML's simplicity for managing Web documents, while exploiting more of SGML's power and flexibility. A large family of applications and related specifications has since emerged around XML. The scope of XML processing and the

complexity of its documentation now surpasses its parent.

Although proprietary formats (like PostScript, PDF, RTF etc.) are still widely in use, there has been an explosion of markup languages and applications based on XML. Today, XML is not only an essential part of the enabling technology underlying the Web, but also plays a crucial role as exchange format in databases, graphics and multimedia applications in sectors ranging from industry, over business and administration, to education and academic research.

Problems addressed

For all the developments in XML since 1998, one thing that has not changed is the understanding of XML documents as serializations of tree structures conforming to the constraints expressed in the document's DTD (Document Type Definition) or some form of schema. This seems very natural, and on our analysis the tight integration of linear form (notation), data structure and constraint language is one important key to XML's success.

Notwithstanding XML's many strengths, there are problem areas which invite further research on some of the fundamental assumptions of XML and the document models associated with it. XML strongly emphasizes and encourages a hierarchical document model, which can be validated using a context-free grammar (or other grammars that encourage a constituent structure interpretation, like context sensitive and regular grammars).

Consequently, it is a challenge to represent in XML anything that does not easily lend itself to representation by context-free or constituent structure grammars, such as overlapping or fragmented elements, and multiple co-existing complete or partial alternative structures or orderings. For the purpose of our work, we call such structures complex structures, and we call documents containing such structures complex documents.

Complex structures are ubiquitous in traditional documents — in printed as well as in manuscript sources. Common examples are associated with the physical organization of the document and the compositional structure of the text, in other words, such things as pages, columns and lines on the one hand, and chapters, sections and sentences on the other. Sentences and direct speech tend to overlap in prose, verse lines and sentences in poetry, speeches and various other phenomena in drama. Complex structures occur frequently also in databases, computer games, hypertext and computer-based literature.

In the last few years problems pertaining to complex structures have received increasing attention, resulting in proposals for

- conventions for tagging complex structures by existing notations, by extending such notations, or by designing entirely new notations;
- alternative data structures;
- explications of the semantic relationships cued by markup in a form that is more easily machine-processable.

The MLCD (Markup Languages for Complex Documents) project aims to integrate such alternative approaches by developing both an alternative notation, a data structure and a constraint language which as far as possible is compatible with and retains the strengths of XML-based markup, yet solves the problems with representation and processing of complex structures.

MLCD started in 2001 and is expected to complete its work in 2007. The project is a collaboration between a group of researchers based at several different institutions. The remainder of this paper presents an interim report from the project

Data Structure

One of the early achievements of MLCD was the specification of the GODDAG (Generalized Ordered-Descendant Directed Acyclic Graph) structure. It was originally based on the realization that overlap (which was the first kind of complex structure we considered) can be represented simply as multiple parentage.

A GODDAG is a directed acyclic graph in which each node is either a leaf node, labeled with a character string, or a nonterminal node, labeled with a generic identifier. Directed arcs connect nonterminal nodes with each other and with leaf nodes. No node dominates another node both directly and indirectly, but any node may be dominated by any number of other nodes.

We distinguish a restricted and a generalized form

of GODDAG. Conventional XML trees satisfy the requirements of generalized as well as restricted GODDAGs. In addition, restricted GODDAGs lend themselves to representation of documents with concurrent hierarchies or arbitrarily overlapping elements, whereas generalized GODDAGs also allow for a convenient representation of documents with multiple roots, with alternate orderings, and discontinuous or fragmented elements.

The similarities between trees and GODDAGs allow similar methods of interpreting the meaning of markup: properties can be inherited from a parent, overridden by a descendant, and so on. There is some chance for conflict and confusion, since with multiple parents, it is possible that different parents have different and incompatible properties.

Recent work has revealed a weakness in the current specification of GODDAG, which leads to problems with the representation of discontinuous elements. In the full version of this paper we will present the results of our work towards a solution to these problems.

Notation

It is always possible to construct GODDAGs from XML documents. In the general case, they will be trees, which are subsets of GODDAGs. It is also possible to construct GODDAGs from the various mechanisms customarily used in order to represent complex structures in XML. However, these mechanisms depend on application-specific processing and vocabularies, and tend to be cumbersome.

Thus, one may either try to establish standards for the representation of complex structures in XML, or provide an alternative notation which lends itself to a more straightforward representation of complex structures. We believe that these options are complimentary, and that both should be pursued.

Thus, we have defined an alternative notation to XML, TexMECS. The basic principles of its design are:

- For documents that exhibit a straightforward hierarchical structure, TexMECS is isomorphic to XML.
- Every TexMECS document is translatable into a GODDAG structure without application-specific processing.
- Every GODDAG structure is representable as a TexMECS document.

A particular advantage of TexMECS is a simple and straightforward notation for what we have called complex structures.

We also plan to design algorithms for translating widely recognized XML conventions for representation of complex structures into GODDAGs, and vice versa. In the full version of this paper we will report on our latest work in this area.

Constraint Language

One of the most important remaining tasks for the MLCD project is the identification of a constraint mechanism which relates to GODDAGs as naturally as constituent structure grammars relate to trees, which constitute a subset of GODDAGs. Constraint languages for XML documents exist in the form of XML DTDs, XML Schema, Relax NG and others. These methods invariably define context-free grammars allowing the representation of XML documents in the form of parse trees. However, since GODDAG structures are directed acyclic graphs more general than trees, they cannot easily be identified with parse trees based on context-free grammars.

Several possible ways forward exist and remain to be explored. MLCD has decided to focus on two approaches, one grammar-based and one predicate-based.

The grammar-based approach starts from the observation that GODDAGs can be projected into sets of tangled trees. One way to achieve at least partial validation of complex documents, therefore, is to write grammars for each such tree and validate each projection against the appropriate grammar. Each such grammar will treat some start- and end-tags in the usual way as brackets surrounding structural units, but treat other start- and end-tags as if they were empty elements. This allows some measure of control over the interaction and overlapping of specific elements in different grammars; whether it provides enough control remains to be explored.

Another approach to validation is to abandon the notion of document grammars, and regard validation simply as the establishment of some set of useful invariants. A schema then takes the form of a set of predicates; the document is valid if and only if all of the required predicates are true in that document. In the XML context,

this approach is represented by Schematron. It is clear that it can also be applied to documents with complex structures, if the language used to formulate the required predicates is extended appropriately.

In the full version of this paper we will report on our attempts to pursue each of these two approaches.

References

Barnard, D.; Burnard, L.; Gaspart, J.; Price, L.; Sperberg-McQueen, C.M. and Varile, G.B. (1995) Hierarchical encoding of text: Technical problems and SGML solutions. *Computers and the Humanities*, 29, 1995.

Barnard, D.; Hayter, R.; Karababa, M; Logan, G. and McFadden, J. (1988) SGML-based markup for literary texts: Two problems and some solutions. *Computers and the Humanities*, 22, 1988.

Dekhtyar, A. and Iacob, I.E. (2005) A framework for management of concurrent XML markup. *Data and Knowledge Engineering*, 52(2):185–215, 2005.

DeRose, S. (2004) Markup overlap: A review and a horse. In *Extreme Markup Languages 2004*, Montreal, 2004. IDEAlliance. Available on the Web at <http://www.mulberrytech.com/Extreme /Proceedings/html/2004/DeRose01 /EML2004DeRose01.html>.

Dubin, D., Sperberg-McQueen, C. M., Renear, A., and Huitfeldt, C. (2003) "A logic programming environment for document semantics and inference", *Literary and Linguistic Computing*, 18.2 (2003): 225–233 (a corrected version of an article that appeared in 18:1 pp. 39-47).

Durusau, P. and O'Donnell, M.B. (2004) *Tabling* the overlap discussion. In *Extreme Markup Languages 2004*, Montreal, 2004. IDEAlliance. Available on the Web at <http://www.mulberrytech.com/Extreme /Proceedings/html/2004/Durusau01 / EML2004Durusau01.html>

Hilbert, M.; Schonefeld, O, and Witt, A. (2005) Making CONCUR work. In *Extreme Markup Languages 2005*, Montreal, 2005. IDEAlliance. Available on the Web at <http://www.mulberrytech.com/Extreme /Proceedings/html/2005/Witt01 /EML2005Witt01.xml>.

Huitfeldt, C. (1995) Multi-dimensional texts in a one-dimensional medium. *Computers and the Humanities*, 28:235–241, 1995.

Huitfeldt, C. (1999) *MECS—A multi-element code system*. Working papers of the Wittgenstein Archives at the University of Bergen. Wittgenstein Archives at the University of Bergen, Bergen, 1999.

Huitfeldt, C. and Sperberg-McQueen, C.M. (2001) Texmecs: An experimental markup meta-language for complex documents. Available on the Web at <http://helmer.aksis.uib.no/claus/mlcd/papers/texmecs.html>, 2001.

Huitfeldt, C. (2003) "Scholarly Text Processing and Future Markup Systems" in Georg Braungart, Karl Eibl, Fotis Jannidis (eds.): *Jahrbuch für Computerphilologie 5 (2003)*, Paderborn: mentis Verlag 2003, pp. 217-233. Available on the Web at <http://computerphilologie.uni-muenchen.de/jg03/huitfeldt.html>

Jaakkola, J. and Kilpeläinen, P. (1998) *Sgrep home page*, 1998. Available on the Web at <http://www.cs.helsinki.fi/u/jjaakkol/sgrep.html>.

Jagadish, H. V.; Laks V. S.; , Scannapieco, M.; Srivastava, D. and Wiwatwattana, N. (2004) Colorful XML: One hierarchy isn't enough. In *Proceedings of the 2004 ACM SIGMOD International conference on management of data, Paris*, New York, 2004. Association for Computing Machinery Special Interest Group on Management of Data, ACM Press

Nicol, G. (2002) Core range algebra: Toward a formal theory of markup. In *Extreme Markup Languages 2002*, Montreal, 2002. IDEAlliance. Available on the Web at <http://www.mulberrytech.com/Extreme / Proceedings/html/2002/Nicol01 /EML2002Nicol01.html>.

Piez, W. (2004) Half-steps toward LMNL. In Proceedings of Extreme Markup Languages 2004, Montreal, Quebec, August 2004.

Renear, A.; Mylonas, E. and Durand, D. (1993) Refining our notion of what text really is: The problem

of overlapping hierarchies. In N. Ide and S. Hockey, editors, *Research in Humanities Computing*, Oxford, 1993. Oxford University Press. Available on the Web at: <http://www.stg.brown.edu /resources/ stg/monographs/ohco.html>.

Sasaki, F. (2004) Secondary information structuring: A methodology for the vertical interrelation of information resources. In *Extreme Markup Languages 2004*, Montreal, 2004. IDEAlliance. Available on the Web at <http://www.mulberrytech. com/Extreme /Proceedings/html/2004/Sasaki01/ EML2004Sasaki01.html>.

Sperberg-McQueen, C.M., and Burnard, L. (eds.) (2001) Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: TEI P4, 2001.

Sperberg-McQueen, C. M., Huitfeldt, C. and Renear, A. (2000) "Meaning and interpretation of markup". *Markup Languages: Theory and Practice 2*, 3 (2000), 215–234.

Sperberg-McQueen, C. M. and Huitfeldt, C. (1999) Concurrent document hierarchies in MECS and SGML. *Literary & Linguistic Computing*, 14(1): 29–42, 1999. Available on the Web at <http://www. w3.org/People/cmsmcq/2000/poddp2000.html>.

Sperberg-McQueen, C. M. and Huitfeldt, C. (2000) Goddag: A data structure for overlapping hierarchies. In P. King and E. Munson, editors, *DDEP-PODDP 2000*, number 2023 in Lecture Notes in Computer Science, pages 139–160, Berlin, 2004. Springer. Available on the Web at <http://www.w3.org/People /cmsmcq/2000/poddp2000.html>.

Tennison, J. and Piez, W. (2002) Lmnl syntax. Available on the Web at <http://www.lmnl.net/prose /syntax/ index.html>, 2002.

Witt, A. (2004) Multiple hierarchies: new aspects of an old solution. In *Extreme Markup Languages 2004*, Montreal, 2004. IDEAlliance. Available on the Web at <http://www.mulberrytech.com/Extreme /Proceedings/html/2004/Witt01 /EML2004Witt01. html>.

Semantic Timeline Tools for History and Criticism

Matt JENSEN
NewsBlip

New concepts in the visualization of time-based events are introduced and applied to the fields of historiography and criticism. These techniques (perpendicular timelines, dynamic confidence links, and time-slice relationship diagrams) extend the semantic power of timelines so that they can show the development of complex concepts and interpretations of underlying events. An interactive software tool called "TimeVis" illustrates these techniques with both 2D and 3D views.

History is a referent discipline. Later events build on earlier events, though in unpredictable and complicated ways. Historiography and literary criticism are the histories of accumulated comments on a subject. The underlying history and literature (the "base events") occur in one era, and commentary and subsequent events ("secondary events") are added later. However, commentary is not made in the same order as the base events; scholars might spend decades analyzing a writer's later works, and subsequently change emphasis to her earlier works.

Visualizing such referent-based relationships through time is very difficult with a single, conventional timeline. The concept of stacked timelines of different eras [Jen03] was introduced to align commentary and consequent events with their referents (Figure 1). This is useful when secondary events are evenly distributed, but less useful when they are concentrated on subsets of the base events. Crossing lines are difficult to interpret, and important early events can end up leading to a forest of arrows. What is more, the x-axes of the two timelines have no relation to each other. This lack of relation is in fact the cause of the criss-crossing lines.

This paper describes three new timeline techniques that can be applied to the study of history, criticism, and other fields with a temporal or referent component. Each technique serves a different research need.

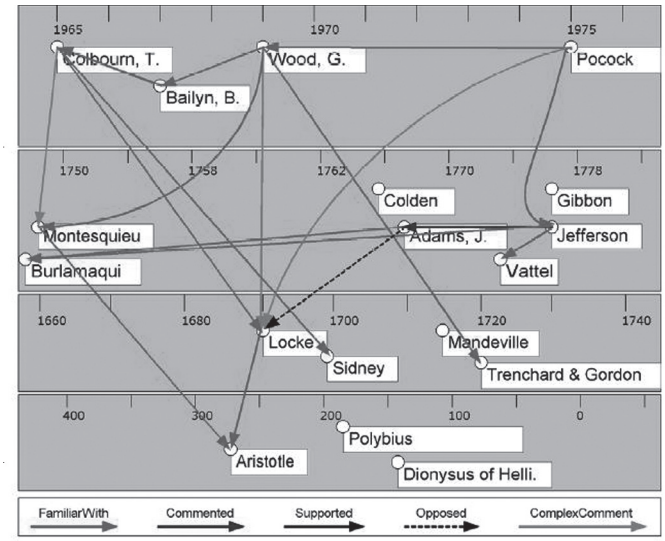


Figure 1. Stacked semantic timeline shows references to earlier timespans.

Techniques

"Perpendicular timelines" are like stacked timelines, but with the second turned 90 degrees, and a second dimension added (Figure 2). The added dimension is the time dimension of the first timeline. This means that secondary events can be visually grouped by the base events that they refer to, yet also be ordered by their own time. In effect, perpendicular timelines allow each original event or topic to spin off its own timeline of commentary and follow-up, arranged perpendicular to the base events.

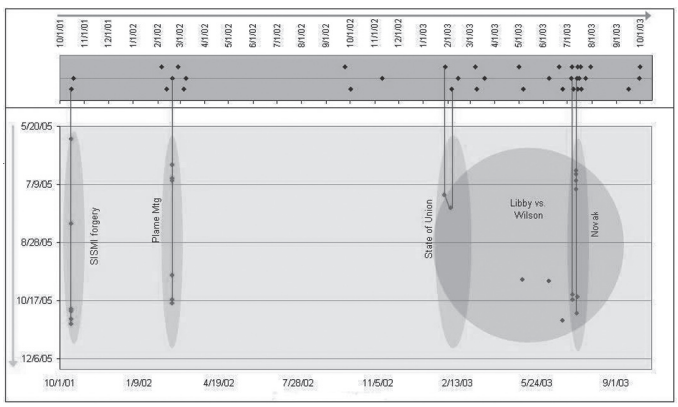


Figure 2. Perpendicular timeline shows subsequent development of older events.

"Dynamic confidence links" build on perpendicular timelines, and provide interactive feedback. Each timeline within TimeVis can have one or more "time slices", which are markers indicating the current point of interest in that timeline (Figure 3).

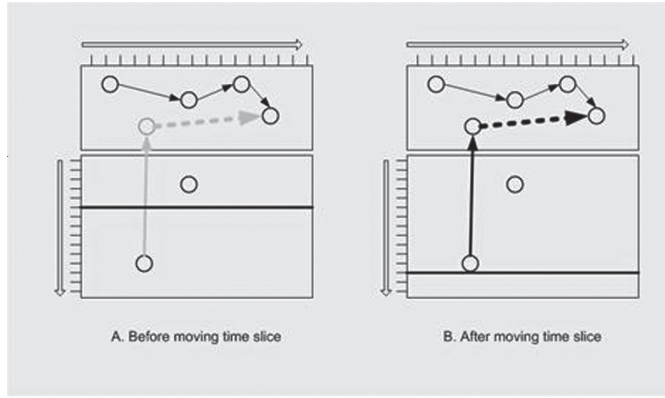


Figure 3. "Dynamic confidence links" are illuminated as the focus in the secondary timeline is shifted (via a slice marker), highlighting revelations of older events.

As we move the current time slice marker in the secondary timeline, we are focusing on "what we knew at time t about the events in the base timeline". As those secondary events occurred in the real world, it changed our interpretation and understanding of the base events. (For example, when Boswell's papers, thought to be lost, were discovered in the early 20th Century, they revealed details of his life that were formerly hidden.) These changed interpretations and understandings can be represented by gray event or concept markers in the base line, connected to the revelatory events of the secondary timeline by gray lines. When the current time slice marker passes a secondary event, its lines and the base events to which it connects can turn from gray to black, indicating that this was the time at which those facts or interpretations became more plausible. That is, the tool dynamically indicates our confidence in different claims, as a function of time.

(Note that the inverse applies as well. If secondary events tend to reduce our confidence in earlier interpretations, those revelations can cause the base events to turn gray.)

Perpendicular timelines, and the dynamic confidence links they enable, can also be extended from two dimensions to three dimensions (Figure 4). Just as the move from one simple timeline to perpendicular timelines frees the secondary events to be organized both by topic and by time, the extension of perpendicular timelines into an

additional dimension allows data to be organized by time as well as two other criteria.

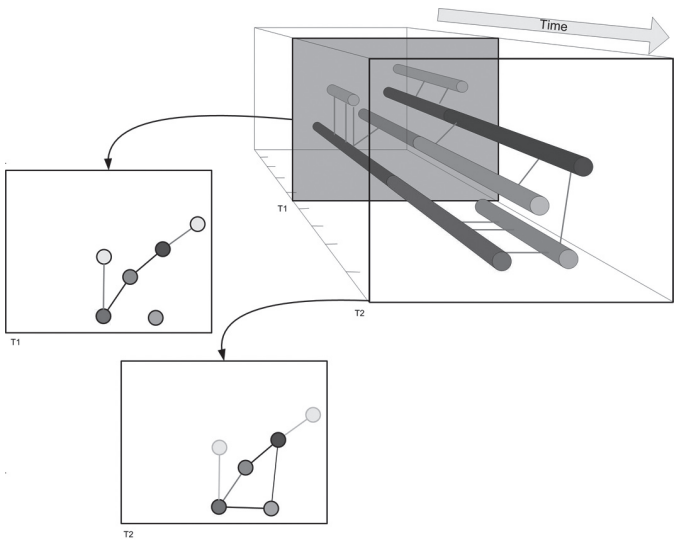


Figure 4. 3D timeline, with two slice markers showing time-slice relationship diagrams. This illustrates the growing importance of the green actor as we compare time T1 to time T2.

Both the base timeline (now a 3D timeline space) and the secondary timeline (also a 3D timeline space) have more flexibility. Rather than use strict definitions for the two free axes in the base 3D space, they can act as a 2D surface for organizing the time-oriented topic/subject bars. By shuffling the time-oriented bars around, the links between them can become more understandable. This is similar to the behavior of force-directed 2D network graph tools such as Visual Thesaurus.

The secondary 3D space can now be organized with one axis for its own events' time, another axis for theme or subject, and the third axis for actor. This allows us to illustrate concepts such as "who knew what when?"

"Time-slice relationship diagrams" show how different actors are related to each other at the time pointed to by the current time slice marker. If events are organized around the concept of actors, and laid out in a 3D space, then a 2D time slice through the 3D space can show a relationship diagram, indicating the "small world" connections from one actor to another. The diagram is equivalent to looking down the time axis of the 3D space, backwards in time, with older connections appearing distant (and thinner) and newer connections appearing large. As the current time slice marker is moved through time, the user can see relationships forming and decaying.

Applications

TimeVis is being used to visualize controversies and cover-ups in history, including Watergate and the Dreyfus Affair. It is being used to investigate the historiographical record of acceptance of Vertot's Roman Revolutions (1719), and the literary response, over the centuries, to Boswell's writings. Those studies should conclude shortly, and while no scholarly breakthroughs are to be expected, what should emerge is a set of visualizations (diagrams, videos, and data files) of use to students and researchers who seek to capture the big picture of such topics that stretch out over time.

References

Akaishi, M. & Okada, Y. (2004), *"Time-tunnel: Visual Analysis Tool for Time-series Numerical Data and Its Aspects as Multimedia Presentation Tool"*, Proc. of the Eighth International Conference on Information Visualization (IV'04).

Chittaro L., Combi C. (2001), *"Representation of Temporal Intervals and Relations: Information Visualization Aspects and their Evaluation"*, Proc. of TIME-01: 8th International Symposium on Temporal Representation and Reasoning, IEEE Press, Los Alamitos, CA, pp. 13-20.

Combi, C., Portoni, L., Pinciroli, F. (1999), *"Visualizing Temporal Clinical Data on the WWW"*, Proc. Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (AIMDM'99), Springer, LNAI 1620, pp. 301-311.

Drucker, J., Nowviskie, B. (2003), "Temporal Modelling", ACH/ALLC 2003.

Jensen, M. (2003), *"Visualizing Complex Semantic Timelines"*, NewsBlip Technical Report 2003001, http://www.newsblip.com/tr/nbtr2003001.pdf .

Knight, B., Ma, J., and Nissan, E. (1998), *"Representing Temporal Knowledge In Legal Discourse, Law,*

Computers, and Artificial Intelligence", Information and Communications Technology Law, Vol.7, No.3 (Special Issue on Formal Models Of Legal Time), pp.199-211.

Konchady, M. (1998), et al, *"A Web based visualization for documents"*, Proc. of the 1998 Workshop on New Paradigms in Information Visualization and Manipulation, ACM PRess, New York, pp. 13-19.

Kullberg, R. (1995), *"Dynamic Timelines: Visualizing Historical Information in Three Dimensions"*, M.S. Thesis at the MIT Media Laboratory. http://robin.www.media.mit.edu/people/robin/thesis/ .

Kumar, V., Furuta, R. (1998), *"Metadata Visualization for Digital Libraries: Interactive Timeline Editing and Review"*, Proceedings of the third ACM Conference on Digital Libraries, p. 126-133.

Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B. (1996), *"LifeLines: Visualizing Personal Histories"*, Proceedings of the CHI '96 Conference on Human Factors in Computing Systems, ACM Press, pp. 221-227.

Richter, H., Brotherton, J., Abowd, G. D., Truong, K. (1999), *"A Multi-Scale Timeline Slider for Stream Visualization and Control"*, GVU Center, Georgia Institute of Technology, Technical Report GIT-GVU-99-30.

Robertson, B. (2000), "The Historical Event Markup and Linking Project", http://www.heml.org.

Rockwell, G. (2003), *"3D Timelines"*, http://strange.mcmaster.ca/~grockwel/weblog/notes/000120.html.

Staley, D. (2002), *"Computers, Visualization, and History"*, AHC Book Series, Sharpe, M.e.

Staley, D. (2000), *"Historical Visualizations"*, Journal of the Association of History and Computing, Vol. III, No. 3, http://mcel.pacificu.edu/JAHC/JAHCIII3/P-REVIEWS/StaleyIII3.HTML .

Tufte, E.R. (2001), *"The Visual Display of Quantitative Information"*, Graphics Press, p.28.

Tominski, C., et al (2005), "*3D Axes-Based Visualizations for Time Series Data"*, Proc. of the Ninth International Conference on Information Visualization (IV'05).

The Inhibition of Geographical Information in Digital Humanities Scholarship

Martyn JESSOP

Centre for Computing in the Humanities
King's College London, Strand
London WC2R 2LS
Phone: 020-7848-2470
Fax: 020-7848-2980
Email: martyn.jessop@kcl.ac.uk

Introduction

Information about place is an essential part of the study of the humanities. People live, events occur, and artefacts are produced by human hand at specific geographical locations and much of what people do is spatially determined or leaves spatial signatures. In order to gain insight into human activity, past and present, the influences of geography must be taken into account. Digital scholarship makes powerful new methodologies freely available at relatively low cost. However, the new research opportunities offered by spatial and spatial-temporal data remain relatively unexplored. This paper examines the reasons for this and discusses possible ways forward for the community.

GIS methodology is much more than digital cartography, it gives the researcher the ability to analyse and display data in a variety of maps, networks or hierarchy trees. The need to represent and model time is leading humanities scholars to experiment with the emerging methodologies of dynamic mapping, an approach that was impossible before the advent of digital scholarship.

There are many ways that methods and tools for structuring, visualising and analysing space, spatial behaviour and spatial change can benefit humanities research. It is over fifteen years since GIS software with reasonable functionality became available in a PC environment at a relatively low cost. Despite this the use of geographical information in digital scholarship by humanists has been poor. This paper will explore some of this potential but,

possibly more importantly, it will also examine why that potential continues to be ignored by many. The author believes there are many reasons, some reflect weaknesses in the methodology and current technologies but possibly the most significant concern our scholarly institutions.

The Developing Role of Geographic information Systems in Humanities research

The use of geographical information in digital humanities research has passed through a sequence of phases of development. Initially the technology was used to replicate pre-existing methodologies and styles of work as in projects such as the Atlas of Mortality in Victorian Britain (Woods and Shelton, 1997). This printed publication used largely standard graphs and cartographic representations to produce an atlas of patterns of health and death in Victorian England and Wales. Later the new methodologies made available by GIS methodology were applied to a variety of new areas for example, the use of 3D digital elevation models to explore the effect that terrain, specifically the gradient required for railway lines, influenced the development of railways in Victorian Britain. Other examples include the dynamic maps used in the Valley of the Shadow and Salem Witch Trial projects. The boundaries of the more rigorously quantitative methods are also being pushed back as the use of 3D and 4D work is explored.

So far the majority of the humanities research performed with GIS has been largely quantitative in nature, but recently there has been increased interest in the use of geographical information for more qualitative work. A number of examples of this style of work can be seen in the Perseus Digital Library, for example the Edwin C Bolles Collection and Boyle Papers. These projects make use of traditional map materials and geographical information in a variety of ways. The Perseus project itself links a number of different digital libraries using geographical information as an integration tool. This work covers a broad range of activities which links texts, images and numerical data to the places they describe on interactive maps to produce an immersive learning environments.

Geographical information also has immense potential both for research and the delivery of information. It provides an unambiguous method for indexing and searching of information. Recent work on map based front-ends to text and image collections has resulted in source discovery tools that are more intuitive and less culturally specific than traditional textual indexing.

Factors Inhibiting the use of Geographical Information in the Digital Humanities

It is clear that geographical information has immense potential both for research and the delivery of information. It provides an unambiguous method for indexing and searching of information offering the ability to build source discovery tools that are more intuitive and less culturally specific that traditional textual indexing. Despite this it has had little effect in the digital humanities.

There area number of limiting factors which need to be addressed

Existing Methodologies: Current GIS methodology is rooted deeply in the origins of the software in the earth sciences. The traditions of rich data sources led to the development of software that is ill-equipped to cope with the sparse and fuzzy data of humanities scholarship. We need to consider how to represent the complexities of the subjects of humanities research visually. What does it mean to think spatially and how do we represent the complex phenomena at work in the humanities visually? Current GIS software has very limited facilities for the handling of time and even these are based on the scientist's view of time as being derived from the phenomenon under study. Humanists require a view of time that is determinative and that can work at different rates and scales moving backwards as well as forwards on a continuous scale.

Humanities Data: Methodologies must be found for tackling issues such sparseness, fuzziness and ambiguity but there are many other broader issues concerning data. The number of digital datasets is growing rapidly and these are often of interest to researchers in fields other than the often highly specialised one that the data was originally derived for but how does one locate them? There is a need for a central archive and a metadata schema that would allow these resources to be discovered. Archivists could provide a community hub for encouraging and implementing the use of spatial content into their retrieval

models thus providing a further means of linking different items of evidence.

Research Practice: Ambitious work with GIS requires a high level of expertise and has a high threshold of usability. The most effective way of facilitating this style of work is through collaboration with researchers and practitioners having backgrounds in fields such as geography, history, information science, computer science, graphic design, and so on. This marks a move away from scholars working independently to a model that necessitates team working which in turn raises further problems. The teams required for this style of collaborative work will be composed of people with very different backgrounds and credentials. There is a strong need to transmit research traditions between disciplines which may have very different agendas and ways of working. There also many issues concerning how the contributions of each member of the team can be acknowledged. For example, for those whose technical work requires considerable expert knowledge far beyond 'technical support' but is fundamentally different from the traditional academic content of the journals and review boards where the published works will be assessed in research assessment exercises such those in the UK. It also raises the issue of how best to prepare students for work in an interdisciplinary team after graduation.

Scholarly Institutions: This innovative style of work is often seen as 'dangerous' and can be seen as posing a career risk for new academics working in environments where successful research assessment is critical. There are limited opportunities for publishing the work and because there are relatively few people with relevant experience at the top of the profession there can be problems with peer review. An additional problem is that projects requiring GIS are often very large and require sustained funding. They also frequently produce resources that will need to be maintained after the completion of the original project. These two funding requirements can pose problems in current research and funding environments.

Scholarly Perception of Geographical Information Science and visualisation: In order to encourage the greater use of geographical information we need a clear statement of the advantages of GIS, and spatial information generally, expressed in terms of research outcomes. This should be supported by a set of exemplar projects and be made available through a 'one stop' source of information. Such a resource could also encourage creative thinking about geographical information and how it can be used in original ways. The fact that humanists are used to working primarily with textual sources may be an inhibiting factor. Many claims are made for the value of GIS as a visualisation tool but it may be that some training is needed in thinking visually and, importantly, the interpretation of the results of visualisation. There could also be more fundamental issues concerning the status and function of images, especially those used for visualization, in humanities scholarship. A similar situation exists with spatial thinking too.

Conclusions

Although we are used to the idea of GIS as a positivist tool its big contribution to the humanities may be as a reflexive one. It can be used to integrate multiple perspectives of the past allowing them to be visualised at various scales. Ultimately it could create a dynamic representation of time and place within culture. This abstract has introduced some of the many factors that are currently limiting the use of geographical information in humanities teaching and research, the final paper will discuss these in more detail and suggest some immediate solutions. The greater use of geographical information could allow us to experience a view of the past that is highly experiential, providing a fusion of qualitative and quantitative information that could be accessed by both naive and knowledgeable alike.

References

Gregory, I., Kemp, K.K., and Mostern, R., 2003. *Geographical Information and Historical Research: Current Progress and Future directions*, Humanities and Computing 13:7-22

Jessop, Martyn (2005). The Application of a Geographical Information System to the Creation of a Cultural Heritage Digital Resource. *Literary & Linguistic Computing* Volume 20 number 1:71-90.

Langren, G., 1992. *Time in Geographic Information Systems*, Taylor and Francis, London and Washington.

Smith, D. M., Crane, G., and Rydberg-Cox, 2000. *The*

Perseus Project: A Digital Library for the Humanities, Literary and Linguistic Computing, 15 (2000), 15-25

Woods, R and Shelton, N., 1997. *An Atlas of Victorian Mortality*, Liverpool.

The SHSSERI Collaborative KnowledgeSpace : a New Approach to Resource Fragmentation and Information Overload

Ian JOHNSON

Archaeology, University of Sydney

Humanities scholars today are faced with information overload created by the explosive growth of resources and services on the web, by globalised instant communication and by the spawning of new personal and intellectual networks outside the confines of discipline and geography. The pace and scale of developments is both exciting and challenging, necessitating the adoption of new strategies which capitalize on the potential of digital methods.

The issues

The explosion in information has not been matched by developments in the conceptual framework and tools we use to manage information, by availability of digital infrastructure or by the widespread adoption of new methods (other than the basic generic tools of email, wordprocessing and web browsing). Where we have adopted digital tools it is often within existing structures – faster communication, easier preparation of publications, easier access to library catalogues and content – rather than in novel approaches to research, teaching or collaboration. Uptake of digital methods in the Humanities is hindered by lack of funding for infrastructure development, by the richness and heterogeneity of our domain, by the lack of agreed methods and classificatory systems and by our relatively slow adoption of technology.

If we are to overcome the problems of information overload we must develop – with very limited resources – e-Research infrastructure and tools which support and enable our tradition of individual scholarship and interpretation, rather than the method- and data-driven

tools of the sciences (or commerce). Our tools must reflect the needs of Humanities scholars – low entry barriers and intuitive structures which reflect the richness and complexity of domain practices. While technologically literate scholars will continue to build or adopt tools and find new ways to use them, how do we enable the rest of the discipline to understand and leverage their potential? The critical need is to develop and propagate authoritative information on Digital Humanities methods repackaged in a form digestible by less technologically-oriented Humanities scholars. We need to get the message out.

SHSSERI Collaborative Knowledge Space

The Sydney Humanities and Social Sciences e-Research Initiative (SHSSERI) is a project to develop e-Research tools and information which reflect the needs of Humanities (and Social Sciences) scholars, particularly tools and which move beyond the mere collection and delivery of digital information into active engagement with research methods and the structure of academic communities. Our aim is to provide a resource which becomes a 'must-have' by providing a single point of access to scattered resources, tools to manage those resources and authoritative information which shortcuts the process of adopting digital tools (and increases the chances of success). Building on existing work at the University of Sydney and using Open Source software, we are developing a one-stop-shop – the SHSSERI Collaborative KnowledgeSpace (CKS) - to support the information, communication, data management, analysis and archiving needs of Humanities and Social Sciences researchers.

In developing the SHSSERI CKS we have identified as our highest priority the management of basic research information (references, bookmarks and notes) and access to authoritative information about the practice and potential of the Digital Humanities. A key focus is the integration of resources. Most researchers store bibliographic references, internet bookmarks and research notes in separate systems (eg. EndNote, Firefox and a Blogg – or on paper). Their CV may be in Word, research observations in Access, research expenditure and grad student details in Excel, email in Outlook, travel arrangements in a corporate system, … the list goes on. Our resources are all too often tied to the desktop of a specific computer, scattered, fragmented, unfindable, inconsistent, redundant, unlinkable and insecure. We work inefficiently, waste time and effort, and stress out trying to manage these disparate, inconsistent resources.

TMBookmarker

The TMBookmarker application (name undecided) tackles the management of bibliographic references, internet bookmarks and note-taking, as well as access to authoritative sources of information on the Digital Humanities. TMBookmarker is a web-accessible knowledgebase which will handle conventional bibliographic information, internet bookmarks and note-taking in a single integrated database. It aims to replace all these forms of machine-specific or special-purpose referencing with a single, integrated searchable database available anywhere one has access to the Internet – from University desk to Internet café in Khatmandu.

In addition to providing consistent anywhere-access and capture/annotation of notes, bookmarks and bibliographic references, the database will generate selective lists for inclusion in web sites, course readings or bibliographies, and allow keywords and annotation to be attached to any resource. It aims to reduce a multiplicity of special-purpose programs, information folders, bloggs, bookmark lists and hand-built web pages to a single, easily understood, web-accessible resource. It will open new avenues of possibility for those who have not ventured into the blogosphere, mastered EndNote or managed to migrate their bookmarks from one machine to the next.

Social bookmarking

TMBookmarker also implements concepts of social bookmarking using a structure we are calling a databliki (database + blogg + wiki). We allow users to discover and share bookmarks/references through the database, while attaching their own personal notes and classifications to them. We provide wiki-based and blogg-based public editing and annotation of references, allowing the community of scholars to participate in expansion and refinement of the reference database, and the development of additional knowledge around the core resource. By mining the database we can identify patterns of bookmarking and knowledge development which link people with similar interests and use the patterns to provide relevancy-based searches of the database.

Unlike generic social bookmarking systems such as del.icio.us or CiteULike, the SHSSERI CKS identifies users by name and institution. This provides a mechanism for identifying colleagues with similar interests and the serendipitous discovery of relevant references through folksonomic tagging tied to specific communities of users. We aim to provide improved folksonomy-based searches by calculating co-occurrence of tags across the database, independent of their actual text values, and using these correlations in combination with user ratings against specific user-defined groups of colleagues and subject domains to generate targeted relevancy measures.

TMBookmarker has many smarts such as instant web bookmarking, automatic DOI identification and lookup, reference import/export from common systems, RSS feeds, saved custom searches, custom list generation and wiki-style change tracking. It has been in use by a test group since Nov 2005 and should be ready for general release by second quarter 2006. It is written in php and MySQL. We intend to release the code as Open Source.

In this presentation I will explore the basic concepts of managing information through a collaborative reference/bookmark/annotation system rather than a conventional desktop reference management system. I will focus on the benefits of integrated access to information in a single database, and the advantages of social bookmarking in an academic referencing system, compared with less structured systems such as del.icio.us or CiteULike. I will also compare the social bookmarking methods developed with those of other academic referencing systems, such as Connotea, and present preliminary results on defining communities of interest through database mining. I will conclude with observations on the relationship between social bookmarking and peer-review methods in establishing relevancy and value of published resources.

Killer Applications in Digital Humanities

Patrick JUOLA

Duquesne University, Dept. of Mathematics and Computer Science

The field of "Digital Humanities" has been plagued by a perceived neglect on the part of the broader humanities community. The community as a whole tends not to be aware of the tools developed by HC practitioners (as documented by the recent surveys by Siemens et al.), and tends not to take seriously many of the results of scholarship obtained by HC methods and tools. This problem has been noticed recently by a number of groups focusing on issues regarding humanities tools, most notably the Text Analysis Developers Alliance (Text Analysis Summit, May 9-11, 2005 at McMasters University) and IATH (Summit on Digital Tools for the Humanities, September 28-30, 2005, at the University of Virginia).

One possible reason for this apparent neglect is a mismatch of expectations between the expected needs of audience (market) for the tools and the community's actual needs. A recent paper by Gibson on the development of an electronic scholarly edition of _Clotel_ may illustrate this. The edition itself is a technical masterpiece, offering, among other things, the ability to compare passages among the various editions and even to track word-by-word changes. However, it is not clear who among Clotel scholars will be interested in using this capacity or this edition; many scholars are happy with their print copies and the capacities print grants (such as scribbling in the margins or reading on a park bench). Furthermore, the nature of the Clotel edition does not lend itself well either to application to other areas or to further extension. It is essentially a service offered to the broader research community in the hope that it will be used, and runs a great risk of becoming simply yet another tool developed by the DH specialists to be ignored. Matthew Jockers has observed that much of the focus in humanities computing is on "methodologies, not results." (Bradley, 2005). This paper argues for a focus on deliverable results in the form of useful solutions to

genuine problems, instead of simply new representations.

The wider question to address, then, is what needs the humanities community has that can be dealt with using HC tools and techniques, or equivalently what incentive humanists have to take up and to use new methods. This can be treated in some respects like the computational quest for the "killer application" -- a need of the user group that can be filled, and by filling it, create an acceptance of that tool and the supporting methods/results. For example, MS-Word and similar software has revolutionized writing, even for non-specialists, to the same degree that Email has revolutionized communication and the World Wide Web has revolutionized publication. They have not only empowered non-specialists to do more, but also created inspiring opportunities for further secondary research in extending the capabilities of these software tools. Digital Humanities needs a "killer application" -- a Great Problem that can both empower and inspire.

Three properties appear to characterize such Great Problems. First, the problem itself must be real, in the sense that other humanists (or the public at large) should be interested in the fruits of its solution. Second, the problem must be interesting, in the sense that the solution is not obvious and the process of solving it will add materially to human knowledge. Third, the problem itself must be such that even a partial solution or an incremental improvement will be useful and/or interesting.

As a historical example of such a Problem (and the development of a solution), consider the issue of resource discovery. With the advent of the Web, billions of resources are now broadly available, but no one knows how to find them. Traditional solutions (journal publications, citation indices, etc.) are no longer adequate as publication can happen through informal channels. Google provides a partial solution to this problem by automatically searching and indexing "the Web," specifically to solve the general problem of finding stuff. At the same time, its algorithms are demonstrably inadequate both in terms of accuracy and in what it can search, leaving much room for incremental development -- but the partial solution that exists has still revolutionized scholarship, and created a huge economic opportunity precisely to extend and improve the solution.

To the three criteria above can thus be added an additional, more political aspect of any proposed "killer app." Any proposed application should be extremely user-friendly, possibly even at the expense of complete generality -- the Perfect should not be the enemy of the Good, especially if the Perfect system is unusably general.

I have argued elsewhere that a possible candidate for such a killer app would be authorship attribution: determining who (if anyone) from a candidate pool of authors wrote a particular document under discussion. This question is obviously of interest, for example, to scholars who wish either to validate a disputed authorship, or for authors wishing to investigate if a document of unknown authorship (say, an unsigned political pamphlet) can be assigned to a known author (and help illuminate some political views). Less obviously, "questioned documents" are often extremely important in a legal and forensic environment -- but traditional forensic analysis, such as handwriting, cannot address questions about (e.g.) born-digital documents, transcriptions, purported copies, and so forth. At the same time, enough papers have been published recently to demonstrate a strong interest in the problem from a humanities standpoint -- and even an analysis that is not strong enough to be conclusive prove can still suggest lines and approaches for further investigation and scholarship.

Another candidate that has been argued elsewhere is automatic back-of-the-book indexing. A third, discussed at the recent IATH summit, is an automatic tool for annotating fully-electronic multimedia documents. These are difficult problems, and a full solution will involve (and illuminate) many subtle aspects of human cognition and of the writing process. At the same time, other scholars will be grateful for the results -- on the one hand, by relieving them of the difficult and expensive burden of generating indices for their own works, and on the other by supporting them in the ability to read and annotate electronic documents in the way they traditionally interact with paper.

It would be to the overall benefit to the DH community to focus at least some effort and resources on the identification and solution of such Great Problems and on the development of such killer apps. The apparent alternative is the status quo, where digital research tools are brilliantly developed, only to languish in neglected disuse by the larger community.

Novel Tools for Creating and Visualizing Metadata for Digital Movie Retrieval

Ilkka JUUSO
Tapio SEPPÄNEN

Department of Electrical and Information Engineering, University of Oulu, Finland

Most of the video information retrieval systems today rely on some set of computationally extracted video and/or audio features, which may be complemented with manually created annotation that is usually either arduous to create or significantly impaired at capturing the content. In this research we set out not only to find the computational features relevant to movies, but also to investigate how much information could be semi-automatically extracted from the production documentation created during the actual production stages a film goes through. It was our hypothesis that this documentation, which has been carefully created for the realization of a movie, would prove highly useful as a source of metadata describing the finished movie.

This paper presents research done at the MediaTeam Oulu research group [1] on identifying and structuring the key metadata descriptors applicable to motion picture content analysis and use of motion picture content in video-on-demand applications. The underlying research includes a study of the concepts involved in narrative structure and a series of empirical viewer tests through which a key set of metadata describing the content and structure of motion picture material was identified. Parts of this research were conducted in co-operation with the Department of Finnish, Information Studies and Logopedics at the University of Oulu [2].

First, established theories and concepts of narration were utilized to examine how movies are structured and what mechanisms are used to package and convey information to the viewer. Studying the formalisms and conventions governing film scripts and screenplays, we attempted to identify and characterize the key elements in movies, e.g. characters, actions and various types of plot points [3]. In addition to these primary elements, we also looked at supporting elements [4], i.e. the kinds of production-related techniques and instruments the filmmakers use in trying to guide the viewer's attention throughout a movie. We found that, for example, editing rhythm and various changes or events in the audiovisual composition of a movie are among the most frequently utilized instruments for highlighting certain sections of a movie where the user should pay more (or less) attention. Our goal in studying these conventions and story-telling instruments was to first understand the domain of the filmmaker - his intended form and function for a movie - before looking at what actions or reactions the movie causes on the part of the viewer.

Secondly, a series of tests using short clips, trailers and an entire movie was carried out in order to investigate how people perceive and process movies [5]. Questionnaires and interviews provided information on what kinds of things viewers notice and how they later describe what they have seen. This information was used to arrive at a key set of metadata that models movies using the same concepts as viewers do in describing them. It was then our task to match these concepts to the instruments used by the filmmakers, in order to find a metadata model that is both usable and feasible to create through a semi-automatic process from what is offered by the movie. Furthermore, the model thus constructed was designed hierarchical in order to facilitate dynamic control over the level of detail in any given metadata category, thereby enabling, for example, the smart summarization of movies on multiple levels. This metadata model then became the starting point for the design of the actual browser that an end-user would utilize in navigating and searching movies.

The next step was to identify the main sources for obtaining this metadata and the best methods for obtaining it. In studying the movie production stages and the documentation relating to each stage, we found that the final script and storyboards, as well as the audio and video tracks of the finished movie were the most interesting and also most practical sources. A comprehensive suite of both automatic and semi-automatic tools for processing these documents and media objects was developed in order to extract the necessary features and information, such as the participants in any given scene, the speakers and the level of interaction between them, motion activity detected on the video

track and a wide range of sound properties extracted from the audio track. These tools included a ScriptTagger, StoryLinker, Scene Hierarchy Editor and a management tool for the underlying database.

The ScriptTagger is a tool which takes in the raw text of the script and turns it automatically into a structured XML document, which can then be further refined semi-automatically to facilitate the tagging of more advanced features. The StoryLinker is a tool used to bring together the script, the storyboards and the video and audio tracks of the movie itself to form individual scenes, where all of the above are linked. The Scene Hierarchy Editor is a semi-automatic tool for grouping together individual scenes using the model constructed on the basis of narrative structure and viewer tests, thus constructing a hierarchical description of the movie. In addition to these, a number of tools were used to extract the audio and video features. A management tool was used to combine the output produced by all the tools above to construct a uniform database.

Ultimately, a prototype of a content-based browser for accessing movies using the metadata model and specialized feature visualization was developed. The implemented browser prototype is a multimodal, content-based retrieval system that enables a wide range of searches based on the hierarchical metadata. This prototype offers users customized views into a movie using user-selectable criteria. The main view offered by the prototype is a navigable hierarchical map of the movie, where the user-selected features are visualized as navigation aids. Users can then move from a higher level down to increasingly detailed descriptions of the movie, i.e. from acts to segments and ultimately to scenes, enabling them to navigate their way down to a particular sequence, ultimately allowing cross comparison of similar sequences in other movies in the database. Alternatively, users may progress through a selected movie on some chosen level of detail, thus enabling them to see the structure of the movie based on the criteria they have chosen. The criteria can be changed on any level or at any point while browsing, i.e. features can be added or removed, as the user sees fit based on how well the features are applicable to the material on any given level and how well they answer the overall search needs of the user. The browser can visualize any new features as long as those features are submitted into the system in the correct format.

The browser and its associated metadata creation tools have numerous applications ranging from, for example, commercial video-on-demand applications for both consumers and media publishing houses to more research oriented ones, such as analysis tools not only for film studies but indeed also for linguistic research into dramatic content, for example of movies and television series. The system could, for example, be used to find certain kinds of conversations or sequences of interest based on their content, structure or audiovisual makeup. The tool is suitable for incorporating non-linguistic information into linguistic information, which has various applications when studying multimodal content, for example in the investigation of expressions of stance, where paralinguistic features complement the linguistic realization of attitude and emotion.

The ideas developed and lessons learned from the construction of these tools and browser will also be applied to a new electronic framework for the collection, management, online display, and exploitation of corpora, which is being developed within the LICHEN project (The Linguistic and Cultural Heritage Electronic Network) [6].

[1] MediaTeam Oulu research group, http://www. mediateam. oulu.fi/?lang=en

[2] Department of Finnish, Information Studies and Logopedics, University of Oulu, http://www.oulu.fi/silo/

[3] Field S. Screenplay – The Foundations of Screenwriting. Dell Publishing. 1994.

[4] Adams B. Mapping the Semantic Landscape of Film: Computational Extraction of Indices through Film Grammar. Curtin University of Technology Ph.D. Thesis, 2003.

[5] Lilja J, Juuso I, Kortelainen T, Seppänen T, Suominen V. Mitä katsoja kertoo elokuvasta – elokuvan sisäisten elementtien tunnistaminen ja sisällönkuvailu. Informaatiotutkimus 23. 2004.

[6] Opas-Hänninen LL, Seppänen T, Juuso I, Hosio M, Marjomaa I, Anderson J. The LInguistic and Cultural Electronic Network (LICHEN): Digitizing and disseminating linguistic data for research and enhancement of the future of minority languages. Second International Conference on Arctic Research Planning (ICARP II), November 10-12, 2005, Copenhagen, Denmark.

Cybercultural Capital: ASCII's Preservation of the Digital Underground

Joel KATELNIKOFF

Department of English & Film Studies, University of Albe

In my conference presentation, "Cybercultural Capital: ASCII's Preservation of the Digital Underground," I will examine independent electronic magazines published in the American Standard Code for Information Interchange, between the years of 1984 and 1993. This period begins with the emergence of the first organized ASCII magazines and ends with the creation of Mosaic, the WWW browser that incorporated HTML and put an end to ASCII's reign as the most widely-used electronic file-type. This nine-year span saw the creation of many independent ASCII magazines, 288 of which can still be accessed through the textfiles.com archives, currently storing over ten thousand issues from the era. These magazines include fiction, poetry, articles, and a plethora of subversive technical manuals on topics such as hacking, virii, and sabotage. Just as Russian Samizdat publishers attempted to undermine the hegemony of the Soviet state through subversive literature, ASCII publishers of North America attempted to undermine Corporate hegemony. In my presentation, I will examine the ruling ethos in ASCII literature, considering cybercultural resistance to corporate paradigms, the cultural need for cyberwriters, and the influence of hacking, sabotage, and computer culture on ASCII fiction and poetry. In an age before the World Wide Web, ASCII text files were a powerful medium for independent publishing, offering disenfranchised suburban cyberpunks easy access to the means of textual production and distribution. While thousands of ASCII texts are currently archived on websites like etext.org and textfiles.com, these websites are maintained by amateurs with no formal training as archivists. As the Internet continues to grow, websites are updated, websites become defunct, and old files are often overwritten by new files. Internet archives are unstable and their documents are at risk of becoming corrupted or erased. My presentation will highlight the literary importance of ASCII texts and explain why an archival project must be undertaken immediately to ensure that the writings of this movement are not entirely lost.

Biography

I have been active in the Canadian independent publishing scene since 1995. In the past decade I have published nearly 400 issues of various ASCII zines, maintained "The Current Text Scene" (a website dedicated to tracking contemporary ASCII zines), and published an ASCII-related article with Broken Pencil. My other areas of interest include creative writing and experimental fiction.

Digital Audio Archives, Computer-Enhanced Transcripts, and New Methods in Sociolinguistic Analysis

Tyler KENDALL
Duke University
North Carolina State University
tsk3@duke.edu

Amanda FRENCH
North Carolina State University
amanda_french@ncsu.edu

Introduction

Traditional methods in sociolinguistic analysis have often relied on the repeated close listening of a set of audio recordings counting the number of times particular linguistic variants occur in lieu of other variants (a classic sociolinguistic example is the tabulating of words using final *–in'* for final *–ing*; cf. Fischer 1958, Trudgill 1974, etc.). These tabulations are normally recorded into a spreadsheet using a program such as Microsoft Excel, or even just into a hard-copy tabulation sheet. The results are then presented as summaries in publications or conference papers as the "data" used for description, explanation, and theory building. Some approaches in linguistics, such as discourse analysis, rely heavily on the development of transcripts of the audio recordings and often the focus of analysis is on the transcript itself and not the original recording or interview event. However, scholars following a wide variety of sociolinguistic approaches have repeatedly highlighted the confounds that arise from these treatments of "pseudo-data" (i.e., analysts' representations of the data) as data. Linguists such as Blake (1997) and Wolfram (e.g., 1993) have discussed problems relating to the tabulation and treatment of linguistic variables and raised the issue that individual scholars' methods are often not comparable. In discussing transcription theory, Edwards has repeatedly pointed out that "transcripts are not unbiased representations of the data" (Edwards 2001: 321). In general, the understanding that linguistic data is more elusive than traditional "hard science" data is widespread but not acted upon. In this paper, we present a project underway at North Carolina State University to argue that computer-enhanced approaches can propel sociolinguistic methodology into a new, more rigorous era.

The North Carolina Sociolinguistic Archive and Analysis Project

The North Carolina Language and Life Project (NCLLP) is a sociolinguistic research initiative at North Carolina State University (NCSU) with one of the largest audio collections of sociolinguistic data on American English in the world. It consists of approximately 1,500 interviews from the late 1960s up to the present, most on analog cassette tape, but some in formats ranging from reel-to-reel tape to digital video. The collection features the interviews of Walt Wolfram, Natalie Schilling-Estes, Erik Thomas, and numerous other scholars. The NCLLP has partnered with the NCSU Libraries on an initiative titled the North Carolina Sociolinguistic Archive and Analysis Project (NC SLAAP). NC SLAAP has two core goals: (1) to preserve the NCLLP's recordings through digitization; and (2) to enable and explore new computer-enhanced techniques for interacting with the collection and for conducting sociolinguistic analysis.

NCSU Libraries has as one of its chief goals the long-term preservation of the recordings made by the NCLLP, and it regards digitization as an appropriate means of preservation. Academic libraries may still be less expert than some commercial organizations when it comes to digitizing and storing audio, but they may be even less equipped to maintain analog audio collections properly (cf. Brylawski 2002, Smith, Allen, and Allen 2004). Archivists and librarians also sometimes point out that digitization and storage of audio may not be worth the expense and difficulty if the sole goal is preservation (cf. Puglia 2003). However, when scholarly digital projects can contribute significantly to the advancement of a discipline, as in the case of NC SLAAP, surely significant investments are called for.

The NC SLAAP project has from the beginning planned to integrate sociolinguistic analysis tools into the archive. This has been achieved to a large degree by integrating

the open source phonetic software application Praat (http://www.praat.org) into the web server software. In brief overview, the NC SLAAP system is an Apache web server currently housed on a Macintosh G5 computer running Mac OS 10.4. Data are stored in a MySQL database and application pages are written in PHP. The web server communicates with third-party open source applications to do most of its "heavy" processing. Most importantly, the web server communicates with Praat to generate real-time phonetic data (such as the pitch data and the spectrogram illustrated in Figure 1).

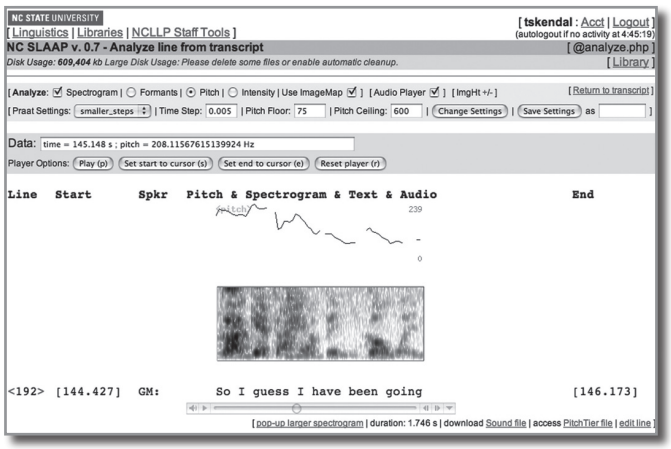


Figure 1: Transcript Line Analysis Example

While certain feature sets are still under development, NC SLAAP, even in its current state, provides a range of tools that greatly enhance the usability of the audio data. These features include an audio player with an annotation tool that allows users to associate notes with particular timestamps, an audio extraction feature that allows users to download and analyze particular segments of audio files, sophisticated transcript display options (as partly illustrated in Figure 1, above), and extensive search and query tools. Importantly, the NC SLAAP software helps to address concerns about the treatment of "pseudo-data" as data, because it enables scholars to better access, check, and re-check their (and their colleagues') variable tabulations, analyses, and conclusions. In short, the NC SLAAP software is an attempt to move us one step – hopefully, a large step – closer to the "real" data.

The features of the NC SLAAP software have potentially tremendous implications for a wide range of linguistic approaches. We focus on only one such feature here: the implications relating to transcription theory.

Transcription Method and Theory

Improvements to the traditional text transcript are extremely important because the transcript is often the chief mediating apparatus between theory and data in language research. Language researchers have long been concerned with the best method and format for transcribing natural speech data (cf. Edwards 2001). Researchers frequently incorporate a number of different transcription conventions depending on their specific research aims. Discourse analysts (e.g., Ochs 1979) traditionally focus most heavily on transcription as theory and practice, but researchers studying language contact phenomena (as in Auer 1998) also have their own transcription conventions for analyzing and presenting their data. At the other end of the spectrum are variationists and dialectologists, who also use transcripts, even if often only for presentation and illustration.

Despite the importance of the transcript for most areas of linguistics, little work has been done to enhance the usability and flexibility of our transcripts. Yet the way a researcher builds a transcript has drastic effects on what can be learned from it (Edwards 2001). Concerns begin with the most basic decision about a transcript: how to lay out the text. Further decisions must be made throughout the transcript-building process, such as decisions about how much non-verbal information to include and how to encode minutiae such as pause-length and utterance overlap. Furthermore, the creation of a transcript is a time- and energy-intensive task, and researchers commonly discover that they must rework their transcripts in mid-project to clarify aspects of the discourse or speech sample.

The NC SLAAP software seeks to improve the linguistic transcript by moving it closer to the actual speech that it ideally represents (Kendall 2005). In the NC SLAAP system, transcript text is treated as annotations on the audio data: transcripts are broken down into utterance-units that are stored in the database and directly tied to the audio file through timestamping of utterance start and end times. Transcript information can be viewed in formats mimicking those of traditional paper transcripts, but can also be displayed in a variety of dynamic ways – from the column-based format discussed by Ochs (1979) to a finer-level focus on an individual utterance complete with phonetic information (as shown in Figure 1, above).

Conclusion

NC SLAAP is a test case for new ways of approaching linguistic analysis, using computers to maintain a strong tie between the core audio data and the analysts' representations of it. In many senses the project is still in a "proof of concept" stage. However, we feel that it has made large steps towards new and more rigorous methods for sociolinguistic analysis and data management. In addition, it can serve as a model for academic libraries as a project that incorporates digital preservation with significant scholarly advancement.

References

Auer, P. (ed.) (1998). *Code-Switching in Conversation*, London: Routledge.

Blake, R. (1997). Defining the Envelope of Linguistic Variation: The Case of "Don't Count" Forms in the Copula Analysis of AAVE. *Language Variation and Change* 9: 57-79.

Brylawski, S. (2002). Preservation of Digitally Recorded Sound, in *Building a National Strategy for Preservation: Issues in Digital Media Archiving*, Council on Library and Information Resources Publication 106.http://www.clir.org/pubs/abstract/pub106abst.html

Edwards, J. (2001). The Transcription of Discourse, in *Handbook of Discourse Analysis*, eds. Deborah Tannen, Deborah Schiffrin, and Heidi Hamilton: 321-348, Oxford and Malden, Ma: Blackwell.

Fischer, J. (1958). Social Influences on the Choice of a Linguistic Variant. *Word* 14: 47-56.

Kendall, T. (2005). Advancing the Utility of the Transcript: A Computer-Enhanced Methodology, paper presented at the Twelfth International Conference on Methods in Dialectology: Moncton, New Brunswick, Canada. August 2005.

Puglia, S. (2003). Overview: Analog vs. Digital for Preservation Reformatting, paper presented at the 18th Annual Preservation Conference, March 27, 2003, at University of Maryland College Park. http://www.archives.gov/preservation/conferences/papers-2003/puglia.html

Smith, A., D. Allen, and K. Allen (2004). *Survey of the State of Audio Collections in Academic Libraries*. Council on Library and Information Resources Publication 128. http://www.clir.org/pubs/abstract/pub128abst.html

Ochs, E. (1979). Transcription as Theory, in *Developmental Pragmatics*, eds. Elinor Ochs and Bambi Schieffelin: 43-72, New York: Academic Press.

Trudgill, P. (1974). *The Social Differentiation of English in Norwich*, Cambridge: CUP.

Wolfram, W. (1993). Identifying and Interpreting Variables. *American Dialect Research*. Ed. Dennis R. Preston. Amsterdam: John Benjamins Publishing Company. 193-221.

How to Annotate Historical Townplans?

Elwin KOSTER

Humanities Computing, Groningen University

In 2003 a project Paper and Virtual Cities, financed by the Netherlands Organization for Scientific Research was started. The ultimate goal is to make explicit choices in the use of historical maps in virtual (re-) constructions for research and design. In order to do so three projects started, focusing on the historical reliability, the technical reliability and on the creation of a mark-up language and visualization tool that could combine the historical and technical data. This presentation is on the toolbox and the mark-up that is created in this third project.

Researchers often use historical maps in order to describe the changes that took place in cities. By using techniques as used in Geographical Information Systems, e.g. rubber sheeting it is possible to make overlays of maps in order to combine data from several sources. Earlier research (Koster 2001) has shown that such combination can provide new insight in the transformation processes in the 17th century townscape. But this process can introduce new errors, for example due to the fact that later copies of maps can depict outdated information. Historical and Architectural Historical research is needed in order to trace this kind of errors.

A second kind of errors is due to the fact that, although the measuring techniques used by 17th century cartographers were very precise, we still need to transform the map in order to make it fit with a modern map. The method of triangulation, measuring angles between high points, e.g. church towers makes that distant points are depicted with high precision, but the area in between might be less accurate. This kind of technical irregularities become visible in the process of rectifying the map. This process is used to bring all the maps to the same scale and orientation so they can be layered and compared.

By storing the technical evidence and the historical evidence in a standardized format, a mark-up language, we are able to connect the two. Such a format visualizes the reliability and veracity of historical town plans and virtual (re-) constructions (separately or in combination) in relation to function and context. This might help an (urban) historian in making choices in studying and describing the change of urban form.

But how would such a standardized format look like? And how can we visualize the reliability and veracity? The extensible markup language (XML) offers different languages that can be used. Historical Events can be stored in Historical Event Markup Language (HEML) while the technical evidence can be captured in the Geography Markup Language. In combination with library standards (Marc-21) used to describe the physical document and RDFPIC to describe the digitized image we are able to describe the document in such a way that data from several maps can be combined into one new standard. In order to do so a new tool has been created that offers the user a way to annotate the digital representation of a map. This tool, with the working title DrawOverMap contains a number of layers in which a researcher, of even a team of researchers can register their data and store in the new markup language.

The data can be used to create a new data layer that visualizes the reliability of the map, or even a combination of maps. With mouse-overs a researcher can query the map, see annotations made by others, and decide what data might be used in research. The user can do so using the DrawOverMap toolbox, but since the XML data is also stored as Scalable Vector Graphics (SVG), again another Markup Language the data can also be browsed interactively over the Internet giving other researchers access to the annotations of historical maps.

With the growing interest in historical maps on the Internet we need a tool to annotate them, rank them and give a valuation on its usability in historical research. We hope that the markup language produced in this research might help historians in choosing which maps to use in their research. On a cultural level this research is important for cultural heritage and for new cultural creations. In the domain of cultural heritage it facilitates the search, the organization and presentation of different sources that are necessary for understanding our urban history. A better understanding of historical developments can lead to an improvement of tools used in town planning.

References

E.A. Koster (2001), Stadsmorfologie (Urban Morphology), Groningen 2001

More on this project:

http://www.let.rug.nl:8080/pvc

HEML: http://www.heml.org

GML: http://www.opengis.org

SVG: http://w3.org/SVG/

Towards the Global Wordnet

Cvetana KRSTEV
Faculty of Philology, University of Belgrade

Svetla KOEVA
Institute for Bulgarian Language - Bulgarian Academy of Sciences

Duško VITAS
Faculty of Mathematics, University of Belgrade

The global wordnet is an extensive lexical-semantic network that constitutes of synonymous sets (synsets) linked with the semantic relations existing between them. The cross-lingual nature of the global wordnet is provided by the establishing of relations of equivalents between synsets that express the same meaning in different languages. The global wordnet offers not only the extensive data for the comparative analysis over lexical densities and levels of lexicalization but furthermore presupposes the successful implementation in different application areas such as cross-lingual information and knowledge management, cross-lingual content management and text data mining, cross-lingual information extraction and retrieval, multilingual summarization, machine translation, etc. Therefore the proper maintaining of the completeness and consistency of the global wordnet is an important prerequisite for any type of text processing to which it is intended.

The EuroWordNet (EWN) extended the Princeton wordnet (PWN) with cross-lingual relations [Vossen, 1999], which were further adopted by BalkaNet (BWN) [Stamou, 2002]. The languages covered by the EWN are Czech, Dutch, Estonian, French, German, Italian, and Spanish, respectively, and those covered by the BWN are Bulgarian, Greek, Romanian, Serbian and Turkish. The equivalent synsets in different languages are linked to the same Inter-Lingual Index (ILI) thus connecting monolingual wordnets in a global lexical-semantic network. The Inter-Lingual Index is based on the PWN (ILI is consecutively synchronized with the PWN versions), the synsets of which are considered as language independent concepts. Thus a distinction between the language-specific modules (English among them) and the

language-independent module (the ILI repository) has to be focused. The ILI is considered as an unstructured list of meanings, where each ILI-record consists of a synset (if the language is not English, a proper translation or at least transliteration must be ensured), an English gloss specifying the meaning and a reference to its source.

Both EWN and BWN adopted the hierarchy of concepts and relations' structure of the English wordnet as a model to be followed in the development of each language-specific wordnet. For the monolingual wordnets a strong rule is observed – strictly to preserve the structure of the PWN because via the ILI a proper cross-lingual navigation is ensured. It is natural, that some of the concepts stored in ILI are not lexicalized in all languages and there are language specific concepts that might have no ILI equivalent. In the first case, the empty synsets were created (called non-lexicalized synsets) in the wordnets for the languages that do not lexicalize the respective concepts. The non-lexicalized synsets preserve the hierarchy and their purpose is to cover the proper cross-lingual relations. Regarding the second case, the ILI is further extended both in EWN and BWN with some language specific concepts. The language specific concepts that are shared between Balkan languages are linked via a BILI (BalkaNet ILI) index [Tufis, 2004]. The initial set of common Balkan specific concepts consisted mainly of concepts reflecting the cultural specifics of the Balkans (family relations, religious objects and practices, traditional food, clothes, occupations, arts, important events, measures, etc).

There are four morpho-semantic relations included in PWN and mirrored in EWN and BWN, Be in state, Derivative, Derived and Participle [Koeva, 2004]. Those relations semantically linked synsets although they can actually be applied to the literals only (graphic and compound lemmas). Consider the following examples:

Be in state is an asymmetric inverse intransitive relation that links derivationally and semantically related adjectives and nouns. The English synset {attractive:3, magnetic:5} with a definition 'having the properties of a magnet; the ability to draw or pull' is in a Be in state relation with the synset {magnetism:1, magnetic attraction:1, magnetic force:1} with a definition 'attraction for iron; associated with electric currents as well as magnets; characterized by fields of force'; also the synset {attractive:1} with a definition 'pleasing to the eye or mind especially through beauty or charm' is in a Be in state relation with {attractiveness:2} denoting 'a beauty that appeals to the senses'.

Derivative is an asymmetric inverse intransitive relation between derivationally and semantically related noun and verb synsets. For example the English synset {rouge:1, paint:3, blusher:2} with a definition 'makeup consisting of a pink or red powder applied to the cheeks' is in Derivative relation with two synsets: {rouge:1} with a meaning 'redden by applying rouge to' and {blush:1, crimson:1, flush:1, redden:1} denoting 'turn red, as if in embarrassment or shame'.

Derived is an asymmetric inverse intransitive relation between derivationally and semantically related adjective and noun synsets. For example the synset {Cuban:1} with a definition 'of or relating to or characteristic of Cuba or the people of Cuba' is in a Derived relation with {Cuba:1, Republic of Cuba:1}.

Participle is an asymmetric inverse intransitive relation between derivationally and semantically related an adjective synset denoting result of an action or process and the verb synset denoting the respective action or process. Consider {produced:1} with a definition 'that is caused by' which is in a Participle relation with {produce:3, bring about:4, give rise:1} denoting 'cause to occur or exist'.

As can be seen by the examples, although the synsets are semantically linked, the actual derivational relations are established between particular literals. For the best performance of the multilingual data base in different text processing tasks a specification of the derivational links must to be kept at the level of literal notes (LNotes).

There are systematic morpho-semantic differences between English and Slavic languages – namely derivational processes for building relative adjectives, gender pairs and diminutives. The Slavic languages possess rich derivational morphology which has to be involved into the strict one-to-one mapping with the ILI.

A vivid derivational process rely Slavic nouns with respective relative adjectives with general meaning 'of or related to the noun'. For example, the Bulgarian relative adjective {стоманен:1} defined as 'of or related to steel' has the Serbian equivalent {čelični:1} with exactly the same definition. Actually in English this relation is expressed by the respective nouns used with an adjectival function (rarely at the derivational level, consider

wooden↔wood, golden↔gold), thus the concepts exist in English and the mirror nodes have to be envisaged.

The gender pairing is systematic phenomenon in Slavic languages that display binary morpho-semantic opposition: male↔female, and as a general rule there is no corresponding concept lexicalized in English. The derivation is applied mainly to nouns expressing professional occupations. For example, Bulgarian synset {преподавател:2, учител:1, инструктор:1} and Serbian synset {predavač:1} that correspond to the English {teacher:1, instructor:1} with a definition: 'a person whose occupation is teaching' have their female gender counterparts {преподавателка, учителка, инструктурка} and {predavačica} with a feasible definition 'a female person whose occupation is teaching'. There are some exceptions where like in English one and the same word is used both for masculine and feminine in Bulgarian and Serbian, for example {президент:1} which corresponds to the English synset {president:3} with a definition: 'the chief executive of a republic', and as a tendency the masculine noun can be used referring to females.

Diminutives are standard derivational class for expressing concepts that relate to small things. The diminutives display a sort of morpho-semantic opposition: big ↔ small, however sometimes they may express an emotional attitude too. Thus the following cases can be found with diminutives: standard relation big ↔ small thing, consider {стол:1} corresponding to English {chair:1} with a meaning 'a seat for one person, with a support for the back' and {столче} with an feasible meaning 'a little seat for one person, with a support for the back'; small thing to which an emotional attitude is expressed. Also, Serbian synset {lutka:1} that corresponds to the English {doll:1, dolly:3} with a meaning 'with a replica of a person, used as a toy' is related to {lutkica} which has both diminutive and hypocoristic meaning. There might be some occasional cases of the expression of that kind of concepts in English, {foal:1} with a definition: 'a young horse', {filly:1} with a definition: 'a young female horse under the age of four', but in general these concepts are expressed by phrases.

There are several possible approaches for covering different lexicalization at different languages [Vitas & Krstev, 2005]:

- treat them as denoting specific concepts and define appropriate synsets;
- include them in the synset with the word they were derived from;
- omit their explicit mentioning, but rather let the flexion-derivation description encompass these phenomena as well.

Treating morpho-semantic relations, relative adjectives, gender pairs and diminutives, in Slavic languages as relations that involve language specific concepts requires an ILI addition for the languages where the concepts are presented (respectively lexical gaps in the rest). This solution takes grounds from the following observations:

- relative adjectives, feminine gender pairs and diminutives denote an unique concept;
- relative adjectives, feminine gender pairs and diminutives are lexicalized with a single word in Bulgarian, Serbian, Czech and other Slavonic languages;
- relative adjectives, feminine gender pairs and diminutives in most of the cases belong to different word class comparing to the word from which they are derived (there are some exceptions, like diminutives that are derived from neuter nouns in Bulgarian) .

Moreover, as with the other morpho-semantic relations, a special attribute assigned at the LNotes must provide information for one-to-one derivational relations.

Although PWN's coverage does not compare yet with new wordnets, the latter are continuously extended and improved so that a balanced global multilingual wordnet is foreseen, thus the task of the proper encoding of different level of lexicalization if different languages is in a great importance regarding the Natural Language Processing.

References

[**Koeva at al.,** 2004] S. Koeva, T. Tinchev and S. Mihov Bulgarian Wordnet-Structure and Validation in: *Romanian Journal of Information Science and Technology*, Volume 7, No. 1-2, 2004: 61-78.

[**Stamou,** 2002] Stamou S., K. Oflazer, K. Pala, D.

Christoudoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou, BALKANET: *A Multilingual Semantic Network for the Balkan Languages, Proceedings of the International Wordnet Conference,* Mysore, India, 21-25 January 2002, 12-14.

[**Tufis,** 2004] D. Tufis, D. Cristea, S. Stamou BalkaNet: *Aims, Methods, Results and Perspectives*.

A General Overview in: *Romanian Journal of Information Science and Technology*, Volume 7, No. 1-2, 2004: 1-32.

[**Vitas & Krstev,** 2005] Duško Vitas, Cvetana Krstev (2005) *Derivational Morphology in an E-Dictionary of Serbian in Proceedings of 2nd Language & Technology Conference,* April 21-23, 2005, Poznań, Poland, ed. Zygmunt Vetulani, pp. 139-143, Wydawnictwo Poznańskie Sp. z o.o., Poznań, 2005.

[**Vossen,** 1999] Vossen P. (ed.) EuroWordNet: *a multilingual database with lexical semantic networks for European Languages*. Kluwer Academic Publishers, Dordrecht. 1999.

Be et *have* : qualités et défauts

Pierre LABROSSE

English Dept, University of Paris-Sorbonne (Paris IV)

Un linguiste français, P. Cotte, fait remarquer que le verbe *be* s'associe avec des adjectifs qualificatifs à connotation positive ou négative (*be courageous, be cowardly*) alors que le verbe *have* est seulement suivi de noms exprimant une qualité (*have courage* vs. * *have cowardice*). La particularité de cette construction *have* + nom est qu'elle donne une caractérisation du sujet avec des noms exprimant une qualité modale focalisée. Contrairement à l'expression de la possession (*John has a car*) où l'objet réfère à une entité différente du sujet, cette qualité est inhérente au sujet. Il semble que le verbe *have* introduise des qualités qui différencient le sujet des autres sujets, dont l'énonciateur. Et s'il n'en introduit pas les défauts, c'est probablement en raison d'une empathie semblable à celle décrite par C. Boisson. L'emploi de noms précédés d'un article zéro indéfini pour exprimer les caractéristiques positives du sujet implique que ces caractéristiques peuvent être partagées par d'autres personnes que le sujet.

En revanche, le verbe *be* peut signaler une certaine distance entre l'énonciateur et le référent du sujet, bien qu'on dise généralement qu'il exprime une identité entre le sujet et son attribut. Cela est probablement dû au fait que son histoire a partie liée avec la deixis. De plus, l'emploi d'adjectifs pour dévoiler les caractéristiques du sujet implique que ces caractéristiques s'appliquent uniquement au sujet, et non à l'énonciateur.

Le linguiste de corpus peut utiliser divers corpus pour vérifier cette hypothèse, même s'il n'est pas possible d'automatiser entièrement la recherche sur ces corpus. Il faut en effet vérifier chaque occurrence pour éliminer des constructions comme *have diesel models* (nom composé pluriel), *Sometimes she had to have water fetched from miles up in the mountain* (structure causative), et, moins fréquemment, … *the emotional strain ambulance men have day in, day out* (circonstants de temps).

J'ai tout d'abord dénombré le nombre d'occurrences

dans le British National Corpus de plusieurs combinaisons pertinentes de *have* + nom (*beauty, courage, intelligence, patience, tact* et leur contraire) ou de *be* + adjectif (*beautiful, courageous, intelligent, patient, tactful* et leur contraire). Le test du χ^2 a révélé une différence significative entre les deux constructions dans tous les cas. J'ai ensuite affiné ce résultat en comparant, pour chaque construction exprimant une caractéristique négative, le pourcentage observé avec le pourcentage théorique. Les résultats ont toujours confirmé l'hypothèse de départ. Mais il n'en va pas de même avec les constructions exprimant une caractéristique positive : dans la majorité des cas, il s'avère que les combinaisons *be* + adjectif sont plus fréquentes que les combinaisons correspondantes en *have* + nom. Ceci montre que l'approche adoptée en linguistique de corpus permet de dégager les constructions qui ne se conforment pas à l'hypothèse, ce qui est un premier pas vers une explication de ces exceptions.

De plus, au fur et à mesure de l'avancement de cette recherche, les listes de concordance mettent en avant de nouvelles caractéristiques des constructions étudiées, comme le caractère [± animé] et [± humain] du sujet. Les premiers résultats tendent à montrer que l'on trouve plus de sujets inanimés avec le verbe *be*.

Nous montrons enfin que le linguiste de corpus peut donner une réponse plus complète à la question posée en interrogeant différents corpus (le BNC et le BNC Baby) avec des logiciels de recherche légèrement différents (SARA ou XAIRA). SARA ne permet pas de formuler des requêtes sur les catégories grammaticales, alors que XAIRA le permet. Cette possibilité, ainsi que le fait que le BNC Baby soit plus petit que le BNC, me permet de trouver quelles qualités ou quels défauts sont le fréquemment exprimés avec les verbes *have* et *be*, ce qu'il serait presque impossible de faire avec le BNC. Les premiers résultats montrent que les qualités et les défauts prototypiques qui intéressent le linguiste théorique ne se rencontrent pas si fréquemment parmi les qualités et les défauts associés à ces deux verbes. Sur les cinq noms de qualité les plus fréquemment associés au verbe *have* (*experience, knowledge, confidence, responsibility, character*), un seul (*character*) est mentionné en linguistique théorique. La situation est encore pire si l'on prend en compte les dix qualités ou défauts associés le plus fréquemment au verbe *be*, c'est-à-dire *able, nice, careful, quiet, honest, difficult, good, bold, important, strong*, en ordre décroissant. Aucun de ces traits de caractère n'est mentionné en linguistique théorique, même si une recherche plus approfondie serait nécessaire pour confirmer ces résultats.

Par conséquent, cette étude en cours montre que le linguiste de corpus peut aborder un problème d'un point de vue assez différent mais toutefois complémentaire à celui d'un linguiste théorique. Celui-ci s'intéresse essentiellement aux concepts ; celui-là s'intéresse à la fois aux concepts et aux occurrences.

Be and *have*: qualities and shortcomings

A French linguist, P. Cotte, remarks that the verb *be* collocates with qualifying adjectives having a positive or negative connotation (*i.e. be courageous, be cowardly*) whereas the verb *have* is followed by nouns expressing a positive quality only (*i.e. have courage* vs. * *have cowardice*). This *have* + Noun construction is quite special in that it expresses a characterisation of the subject with nouns expressing modal qualities which are focalised. Contrary to the usual expression of possession (*John has a car*) in which the object refers to an entity different from the subject, these qualities are part and parcel of the subject. The verb *have* seems to introduce qualities which differentiate this subject from other subjects, and among them, the speaker. But it does not introduce his/her shortcomings most probably because of an empathy akin to that delineated by C. Boisson. The use of nouns preceded by a zero determiner to express the positive characteristics of the subject implies that these characteristics can be shared by people other than the subject.

On the other hand, the verb *be* can somehow distance the speaker from the referent of the subject although it is said to express an identity between the subject and its complement. This is probably due to the fact that its history is linked to deixis. The use of adjectives to express the characteristics of the subject implies that these characteristics apply to the subject only, not to the

speaker.

The corpus linguist can use various corpora to verify this hypothesis even if research on these corpora cannot be fully automated. One needs to check each and every hit to eliminate constructions such as *have diesel models* (plural compound nouns), *Sometimes she had to have water fetched from miles up in the mountain* (causative constructions) and, less frequently, … *the emotional strain ambulance men have day in, day out* (time adverbials). Then one can make the proper calculations and use the appropriate statistical tools.

First I counted how many occurrences of several relevant combinations of *have* + Noun (*beauty, courage, intelligence, patience, tact* and their opposites) or *be* + Adjective (*beautiful, courageous, intelligent, patient, tactful* and their opposites) there were in the British National Corpus. The χ^2 test nearly always showed a significant difference between the two constructions. To refine this result I compared the percentage observed for each construction expressing a negative characteristic with its theoretical percentage (variance test). The results always confirmed the hypothesis. But it was not the same with constructions expressing a positive characteristic: a majority of *be* + Adjective combinations proved to be more frequent than the corresponding *have* + Noun combinations. This shows that a corpus linguistics approach helps isolate the constructions that do not comply with the hypothesis, which is a first step towards finding an explanation to these exceptions.

Moreover, as the research progresses, the concordance lists highlight new characteristics of the constructions under study, such as the [± animate] and [± human] features of the subject. Preliminary results tend to show that there are significantly more inanimate subjects with the verb *be*.

Finally, we see that searching through different corpora (the BNC and the BNC Baby) with slightly different software programmes (SARA or XAIRA) enables the corpus linguist to give a more complete answer to the question raised. SARA does not answer queries about grammatical categories while XAIRA can. This, together with the fact that the BNC Baby is much smaller than the BNC, allows me to find out what qualities or shortcomings are more frequently expressed with the verbs *be* or *have*, something that would be almost impossible to do with the BNC. Preliminary results show that the typical qualities and shortcomings in which the theoretical linguist is interested do not appear so frequently in the qualities and shortcomings associated with the verbs *be* and *have*. Among the five most frequent quality nouns associated with *have* (*experience, knowledge, confidence, responsibility, character*), only one (*character*) is mentioned by the theoretical linguist. The situation is even worse if one considers the ten most frequent qualities or shortcomings associated with *be*, which are, in decreasing order: *able, nice, careful, quiet, honest, difficult, good, bold, important, strong*. None of those traits is mentioned by the theoretical linguist, although further research is necessary to confirm those findings.

This ongoing study thus shows that a corpus linguist can consider the problem from a point of view which is quite different but complementary to that of the theoretical linguist: the latter is mainly interested in concepts; the former both in concepts and occurrences.

References

Boisson, C. (1987). " Anglais HAVE, français AVOIR et l'empathie ". in *La Transitivité*. Saint Etienne : C.I.E.R.E.C. Travaux 52. pp. 155-181.

Cotte, P. (1996). *L'explication grammaticale de textes anglais*. Paris : Presses Universitaires de France. Coll. Perspectives anglo-saxonnes.

Cotte, P. (1998). " *Have* n'est pas un verbe d'action : l'hypothèse de la réélaboration ". *in* Rousseau, A. (éd.) (1998). pp. 415-439

Rousseau, A. (1998). *La transitivité*. Villeneuve d'Ascq : Presses Universitaires du Septentrion. Coll. ULS Travaux et Recherches.

Constructing the Catalogue of English Literary Manuscripts 1450 - 1700

John LAVAGNINO

Centre for Computing in the Humanities, King's College London

Like many people all over the world today, I'm involved in a project to transform a scholarly work from book form to online form. There is no longer any serious doubt about the value of such efforts, especially when the book in question is a reference work; but everyday scholarly experience shows that there are many questions about what sort of online form is best, and in a few cases reviewers have written analyses that detail the obstacles that can exist to the most advanced applications (see, for example, Needham, "Counting Incunables"; Daunton; Williams and Baker). One lesson of such reviews is how different the nature and uses of our various reference works are. All the same, some reflections on the general questions involved can be productive; this talk seeks to cover the relatively familiar issues of searching, and the less frequently discussed question of results display and organization.

Our book is *the Index of English Literary Manuscripts* covering the period from 1450 to 1700, originally compiled by Peter Beal and published in four volumes from 1980 to 1993. Its aim was to catalogue all known literary manuscripts of a selection of writers; it was organized by author and work, not by the contents of manuscripts as is the norm. Those working on more modern material expect to encounter manuscripts that are in an author's own hand; but most of those in the Index are copies by other hands, often combined in miscellanies with works of many other authors. Apart from facilitating work on the individual authors who were covered, the Index also spurred work on the nature of textual transmission in early modern Britain, where scribal publication continued to be important despite the advent of print (see Love).

In making the case for the value of an online version of such a work, a standard claim is to point to the improved access that can be provided. It is less expensive to distribute the completed work online than as a set of hefty books, and it is also easier to find certain kinds of information in a searchable online publication. But as we know from the World Wide Web, even badly-done and inaccurate online resources can be put to use, so we need to look to some goal beyond this absolutely minimal one. As the analyses I've cited by Needham and others show, many uses that scholars can imagine for indexes and catalogues turn out not to be supported in online versions.

In their transformation into online resources, many books go through a process of atomization into individual items, which may then be reunited in various ways: so that information once accessible only through the sequential order of the book, or through manually constructed aids such as indexes, might now be available in many ways. But, of course, these alternative routes depend on the data and on the machinery used to work with it: problems searches involving difficult forms of information such as dates are by now very familiar, and they stem from inconsistency of practice, lack of sophistication in search machinery, and problems in dealing appropriately with uncertainty.

In our case the information was compiled not only following the usual sort of guidelines, but by only one person, so there is at least some chance of doing a reasonable job on this score. A further question is how well we can provide not just search results but an orderly view of a different perspective on the information. By this I mean a display that is *organized and focused* following the desired point of view, rather than merely being a set of search results. We are familiar with the way that a search on minimally-structured full-text resources produces a set of results that need some working through: the scope of the resulting passages is often not clearly delimited and you need to read around to figure out for yourself what the relevant piece of text is. The sequence of such results is also approximate: experienced scholars learn not to pay too much attention to it. These are systems that do not focus on exactly the results called for, and do not organize the results in the best manner; and improving on these matters in a full-text system is difficult. In a catalogue, or anything else that is based on more readily atomized information, there should be more scope for building new perspectives and not just lists of results, but this is another problem that has often proven difficult to solve (see Needham, "Copy Description"). This is more

than a mere question of agreeable presentation: as work in the field of visualization has shown, finding ways to present known data such that our minds can work on it is a powerful method, and one that we need more of in the humanities.

The printed Index of English Literary Manuscripts was an extreme example of a resource that offered only one perspective on information of interest from many perspectives, since there was never an index to offer alternative ways into the information. As examples will show, though, it is not always straightforward to build a display from a new perspective: in our case we are trying to preserve the original author/work perspective but also offer an organization by manuscript. That calls for more than just a reordering of entries: we find that the material within entries needs (at the least) to be rearranged, because even its organization expresses a perspective on the material. There may be a limit to the flexibility of this sort of catalogue, but an awareness of the issue when it's being built can help us improve the design.

Works Cited

Beal, Peter, compiler, *Index of English Literary Manuscripts* (London: Mansell, 1980–1993), 4 volumes.

Daunton, Martin, "Virtual Representation: *The History of Parliament on CD-ROM*", Past and Present 167, May 2000, 238–261.

Love, Harold, *Scribal Publication in Seventeenth-Century England* (Oxford: Clarendon Press, 1993).

Needham, Paul, "Copy Description in Incunable Catalogues", *Papers of the Bibliographical Society of America* 95:2, June 2001, 173–239.

Needham, Paul, "Counting Incunables: the IISTC CD-ROM", *Huntington Library Quarterly* 61:3–4, 1998, 457–529.

Williams, William Proctor, and William Baker, "Caveat Lector. English Books 1475–1700 and the Electronic Age", *Analytical and Enumerative Bibliography* NS 12:1, 2001, 1–29.

Les technologies de l'information et de la communication (TIC) aux services de l'enseignement de la logique aux apprenants des disciplines d'humanités : LOGIC, un outil en ligne, dynamique et interactif.

Florence LE PRIOL
Jean-Pierre DESCLÉS
Brahim DJIOUA
Carine LE KIEN VAN

LaLICC, Université Paris-Sorbonne/CNRS

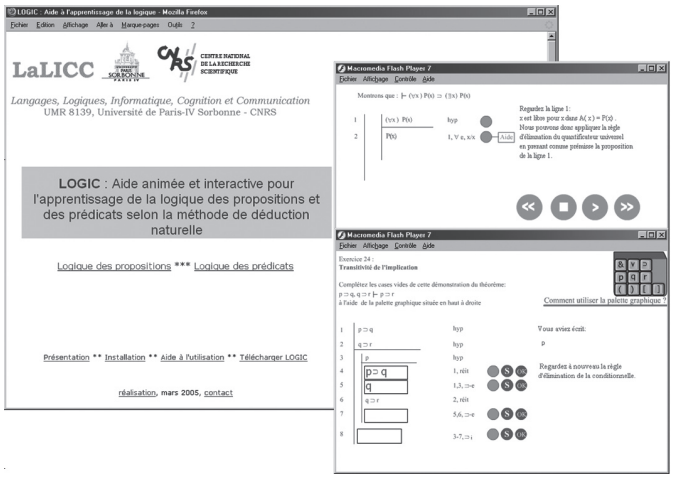
LOGIC is a tool which brings, with learning from the disciplines of humanities, a complement with the lecture of logic, based on the method of natural deduction of Gentzen, exempted by the teacher, in their propose of the dynamic examples and the interactive exercises. It is a free tool, available on line (http://www.lalic.paris4.sorbonne.fr/LOGIC/).

Se servir de l'informatique comme un outil dynamique et interactif : dynamique en montrant comment se déroule, dans le temps, une démonstration ; interactif en faisant intervenir l'apprenant qui est ainsi sollicité de façon active, pour compléter les raisonnements. En cas de difficultés, l'apprenant peut demander en ligne une aide et obtenir la solution locale et une indication des connaissances utiles à la résolution du problème posé.

Ainsi, on passe, grâce au logiciel, d'une lecture passive d'un ouvrage à une appropriation active d'un savoir par une série ordonnée d'exercices où l'apprenant est sollicité en lui donnant, à chaque pas, les possibilités d'être secouru. On voit donc par ce logiciel, qui peut être appelé à tout instant, l'apport de l'informatique à un enseignement

interactif et actif.

LOGIC est un outil qui apporte, aux apprenants des disciplines d'humanités, un complément au cours magistral de logique dispensé par l'enseignant, en leur proposant des exemples dynamiques et des exercices interactifs.



La logique est l'art de bien raisonner, la discipline de la déduction, des démonstrations rigoureuses, de la mécanisation des preuves…

Mais la logique est aussi le lieu des interprétations, de la signification des énoncés, celui des modèles ou mondes possibles.

Ainsi, la logique se construit dans l'opposition entre syntaxe et sémantique : la syntaxe est le monde des symboles, des opérations grammaticales vides de tout contenu, la sémantique est le lieu des interprétations, des modèles ou mondes possibles, le lieu des réalisations, le lieu où une signification est donnée.

En logique mathématique, on doit distinguer entre une conception "axiomatique" de la logique, qui fût celle de Frege, Russel et Hilbert, et une conception plus "pragmatique" en terme d'actes de preuves, que l'on retrouve dans les systèmes de déduction naturelle de Gentzen.

LOGIC est basé sur la méthode de déduction naturelle de Gentzen qui présente la notion de démonstration de manière tout à fait naturelle : par exemple, elle tend à imiter la manière spontanée du mathématicien ; elle permet de montrer comment certaines preuves déductives propagent l'évidence ; elle tend à expliquer le sens des symboles logiques pris isolément.

Prenons le raisonnement suivant, exprimé en langue naturelle :

(1) *Si le ciel se couvre, il risque de pleuvoir. S'il risque de pleuvoir, il est bon de prendre un parapluie. Donc, si le ciel se couvre, il est bon de prendre un parapluie.*

Posons les abréviations suivantes pour simplifier les écritures :

p = *le ciel se couvre*

q = *il risque de pleuvoir*

r = *il est bon de prendre un parapluie*

Le raisonnement (1) s'exprime par l'expression (2), avec les propositions élémentaires p, q et r :

(2) si p, q ; si q, r ; donc si p, r

Nous avons un raisonnement où la dernière proposition (si p, r) est déduite des deux premières (si p, q et si q, r). Introduisons le connecteur propositionnel $\supset$ et le connecteur de conjonction &. L'expression (2) du raisonnement s'exprime maintenant par (3) :

(3) $(p \supset q)$ & $(q \supset r)$ donc $(p \supset r)$

Exprimons maintenant le donc déductif par :

Si les hypothèses $(p \supset q)$ et $(q \supset r)$ sont posées,

Alors il s'ensuit que l'on a $(p \supset r)$

Nous obtenons l'expression (4) :

(4) si $(p \supset q)$ & $(q \supset r)$ donc $(p \supset r)$

Pour mieux exprimer le rôle des hypothèses et celui de la conclusion, nous exprimons l'expression (4) par une déduction naturelle :

1 $p \supset q$ hyp
2 $q \supset r$ hyp
3 $p \supset r$ 1, 2

Nous indiquons clairement les hypothèses aux lignes 1 et 2. Nous indiquons la conclusion à la ligne 3, en mentionnant les lignes qui sont les prémisses de la conclusion.

Comment pouvons-nous affirmer que la conclusion

exprimée à la ligne 3 est déduite des hypothèses exprimées aux lignes 1 et 2 ? Quelles sont les règles qui justifient une telle décision ? La méthode de déduction naturelle précise justement les règles et l'utilisation des règles qui permettent de justifier certains raisonnements (les raisonnements valides) et de rejeter les autres raisonnements.

Une déduction se présente comme une suite de lignes où :

- chaque ligne est identifiée par un numéro (le numéro de la séquence dans la déduction) ;
- chaque ligne exprime une proposition qui est soit posée comme hypothèse, soit déduite des lignes précédentes en appliquant les règles d'élimination ou d'introduction ;
- chaque ligne se termine par une justification qui indique
  - soit la (ou les) règle(s) utilisée(s) et les prémisses appelées par la (ou les) règle(s), ces prémisses étant identifiées par leurs numéros séquentiels ;
  - soit le statut d'hypothèse de la proposition exprimée à cette ligne.

D'une façon générale, une déduction est une suite de propositions $P_1, \ldots, P_i, \ldots, P_n$ où chaque proposition P_i est soit une hypothèse que l'on introduit, soit une conclusion déduite des prémisses $P_1, \ldots, P_{i-1}$ déjà déduites ou introduites comme des hypothèses.

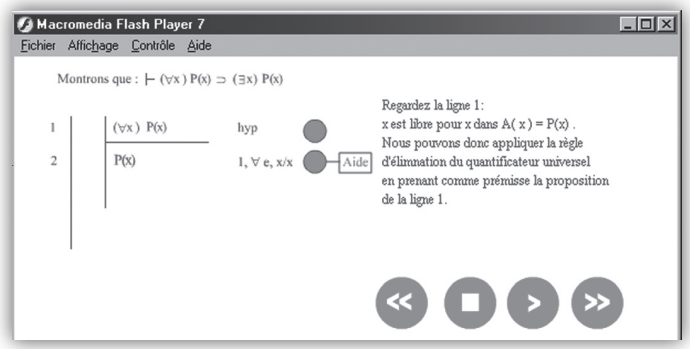
Les schémas de règles de la méthode de déduction naturelle sont soit des schémas d'élimination, soit des schémas d'introduction d'un symbole logique.

LOGIC propose aux apprenants, à travers une interface web (http://www.lalic.paris4.sorbonne.fr/LOGIC/) compatible avec tous les navigateurs acceptant le lecteur d'applications Flash, un rappel des principaux éléments théoriques (en particulier les règles d'introduction et d'élimination des opérateurs) illustrés par des exemples dynamiques et des exercices interactifs de trois catégories.

Le cours est organisé en deux parties. La première partie est consacrée à la logique des propositions et comprend 9 exemples et 64 exercices répartis en quatre blocs et deux séries d'exercices formels en langue naturelle. La seconde partie est consacrée à la logique des prédicats et comprend 7 exemples, 46 exercices répartis en neuf blocs. Ce découpage permet d'introduire les notions les unes après les autres et de grouper les exercices par règles, donnant la possibilité aux apprenants de réaliser un parcours d'apprentissage progressif.

Dans chaque bloc introduisant des règles, le fonctionnement de chaque règle est illustré par des exemples dynamiques.

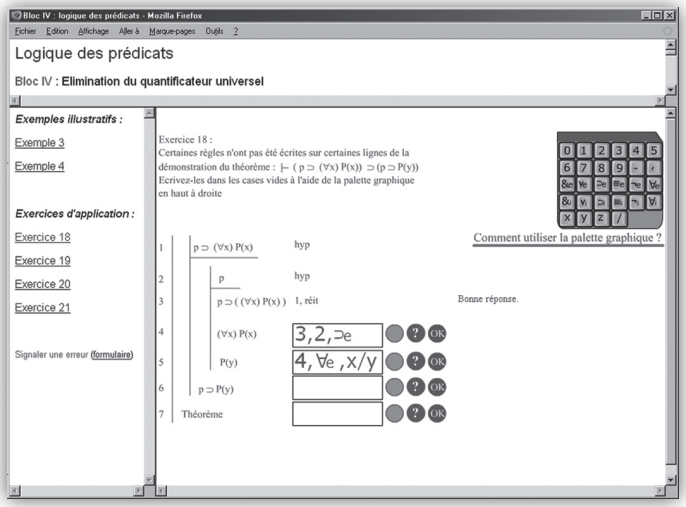
L'animation des exemples présente de nombreux avantages par rapport à une version statique (sur papier, par exemple), en effet, elle reproduit la présentation que pourrait en faire le professeur au tableau, étape par étape. Elle présente également des avantages sur la présentation du professeur car chaque apprenant peut aller à son rythme et reprendre à sa guise les points qui lui pose problèmes. Une fois démarrée, l'animation montre comment la déduction est construite ligne par ligne en modifiant la couleur des propositions mise en jeu dans la règle en cours d'exécution et en construisant la nouvelle ligne dynamiquement. Dans chaque animation, des boutons permettent soit d'obtenir de l'aide, soit de modifier le déroulement de l'animation.



Ecran des exemples animés

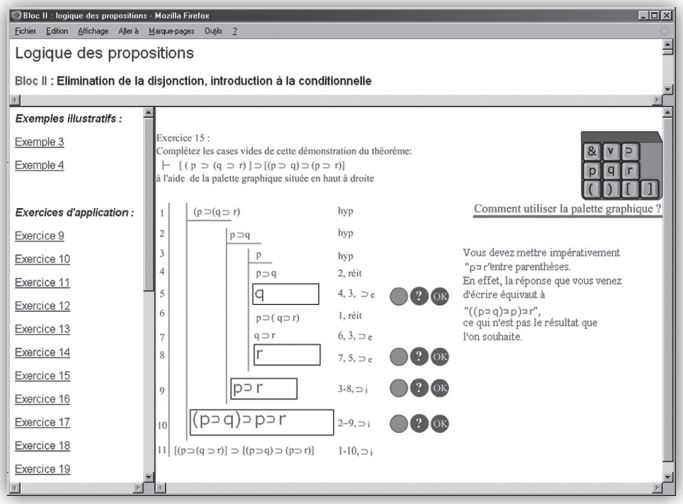
Les exercices interactifs sont proposés aux apprenants afin de leur permettre de tester leur compréhension du problème. Ces exercices peuvent être classés en trois catégories :

- les exercices où certaines règles utilisées lors de la déduction doivent être trouvées : dans ces exercices, certaines règles qui ont été appliquées lors de la déduction n'ont pas été indiquées, l'exercice consiste donc à étudier les expressions afin de déterminer la règle qui a été utilisée ;



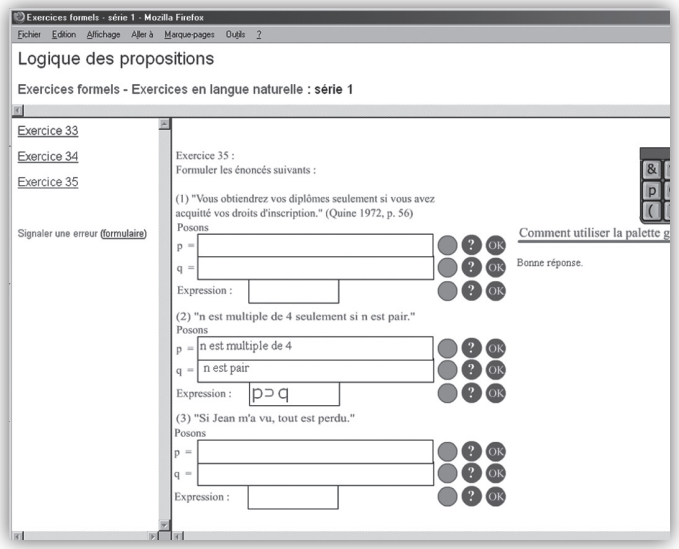
Exercice où déterminer les règles

- les exercices où certaines expressions de la déduction manquent : dans ces exercices, les règles qui ont été appliquées lors de la déduction sont toutes écrites mais certaines expressions manquent, l'exercice consiste donc à appliquer la règle pour déterminer l'expression ;



Exercice où déterminer les expressions

- les exercices basés sur le langage naturel : dans ces exercices, il faut formaliser l'énoncé en langue naturelle dans la logique des propositions ou dans la logique des prédicats et parfois, faire la démonstration.



Exercice formel

L'apprenant déroule les exercices à son rythme, il écrit et valide ses réponses ou peut demander de l'aide ou la solution.

LOGIC est utilisé notamment par les étudiants de l'université Paris-Sorbonne en Licence « Lettres classiques et modernes, sciences du langage », parcours « Langue Française et Techniques Informatiques » et en Master « Information et Communication. Informatique et Ingénierie de la Langue pour la Gestion de l'Information », parcours « Logique, Sémantique, Cognition et Informatique ».

Ce type d'outil, qui permet à l'apprenant d'approfondir la formalisation logique des énoncés en langue naturelle en logique des propositions ou en logique des prédicats, est encore assez peu fréquent sur internet, en langue française.

Références

Desclés, J-P. (1995). *Méthode de la déduction naturelle (d'après Gentzen)*. Cours du DEA MIASH. Université Paris-Sorbonne

Gentzen, G. (1934). *Untersuchungen über das logische Schliessen*. Mathematische Zeitschrift 39

Le Kien Van, C. (2002). *Cours interactif de logique*.

Mémoire de DEA. Université Paris-Sorbonne

Desclés, J-P., Djioua, B., Le Kien Van, C., Le Priol, F. (2005). *LOGIC : Aide dynamique et interactive pour l'apprentissage de la logique des propositions et des prédicats selon la méthode de déduction naturelle*. Cahiers LaLICC n°2005/01

Our Daily Dose of Rhetoric : Patterns of Argument in the Australian Press

Carolyne LEE

Media & Communications Department, University of Melbourn

As twenty-first century Australian citizens, the main way in which we "do" democracy is through the media. Every important issue in public life comes to us via the mass media, almost always structured as some form of 'argument'. It is therefore important that these arguments are not only analysed, dissected and evaluated, but are also presented in ways that enable as wide a section of the population as possible to understand them, so that when citizens engage in deliberation about public issues, their dialogue will be logical, systematic and dialectical. Dialectic is two-sided. It is argument that is constructive and balanced. Now more than ever, at a time when "social, cultural and religious issues are of growing significance due to the insecurities of globalization and the increasing role of non-state players in the security environment" (Aust. Govt. 2004-2005: 9), we need not only greater scrutiny of mediatised arguments, but also stronger strategies for "[E]nhancing our nation's understanding of social, political and cultural issues [which] will help Australia to engage with our neighbours and the wider global community and to respond to emerging issues" (Aust. Govt.: 8)

To this end, my aims are threefold: to analyse a range of important public issues presented as arguments in the newspapers read by the majority of Australians, using the argument-diagramming computer software *Araucaria* (Reed and Rowe 2001); to evaluate each argument thus represented as diagrams, assessing each one in terms of its logical and dialectic reasoning; lastly, to use the diagrams to develop strategies for maximizing public understanding of important issues. A discussion of a small section of the research leading to the first two aims forms the content of the current paper, research that involves the trialling of the conjunction of argumentation schemes (Walton 1996) with the software Araucaria (Reed and

Rowe 2001), capable of diagramming the construction of reasoning in a particular set of argumentative texts. The texts analysed are two sets of 'columns' of argumentative journalism ('op-eds') from the two 'rival' daily newspapers in Melbourne Australia, *The Age*, and the *Herald Sun* (the readership of both together comprising the largest per-capita newspaper reading population in Australia), owned respectively by Fairfax Ltd., and Rupert Murdoch's News Ltd. Research of this type has not been done before in Australia in relation to newspaper arguments. Despite many studies on propaganda, bias, ethics, and news framing in the media, often focusing on the Australian media (see for example Hirst and Patching 2005), systematic empirical research into argumentative structures of news writing is tantalizingly absent.

Not only is it important that mediated arguments be subject to analysis and critique, but also that the findings from such research be presented so that a wide section of the public can understand logical argument structures, and thus learn to decide for themselves whether arguments in the media are logically and dialectically reasoned. This is turn will aid citizens to engage optimally in deliberation about important public issues. Research suggests that currently this is not occurring (Kuhn 1991, Van Gelder 2003). As Van Gelder states: "Deliberation is a form of thinking in which we decide where we stand on some claims in the light of the relevant arguments. It is common and important, whether in our personal, public or working lives. *It is also complicated, difficult and usually poorly done*" (Van Gelder 2003, emphasis added). My project seeks to address this problem by presenting some of the results in computer-generated graphic format—by making argument maps. In this way, the inconsistencies, biases, and other fallacies will be able to be clearly understood by as wide a section of the public as possible. The benefit of such maps over verbal explanations has been articulated by Van Gelder (2003): "Although prose is the standard way to present reasoning, it is not a good tool for the job. Extracting the structure of evidential reasoning as typically presented in prose is very difficult and most of the time we do it very badly…" It is hoped that my project will serve to make argument-mapping more widely known and understood, in terms of both its use in assessing arguments presented in the media, as well as in its potential to improve reasoning and deliberation in the public sphere.

It is likely that some or many arguments will be found—by way of the argument-diagramming and subsequent evaluation—to be one-sided or dialectically ill-reasoned. The proportion found to be like this will have a major bearing on the conclusions of my project. If, for example, the majority of the arguments are found to be invalid, and not to contain a sequence of reasoning, this has serious implications for public understanding of and reasoning about important issues—that is, the skills of deliberation. Some research has suggested that these skills are not possessed by up to half the population (Kuhn 1991). But is this inability to deliberate over public arguments due to lack of "basic reasoning and argument skills", or due to arguments presented by the media in ways that are illogical, faultily-reasoned or one-sided? Is it possible that flawed or fallacious arguments in the public sphere might in some way be causally linked to this lack of basic reasoning and argument skills, and so to the inability of many to deliberate over important issues? Such questions underpin the aims of my project, because deliberation is crucial in any public sphere, not simply because it is about achieving consensus, but rather because it about seeking knowledge.

References

Australian Government: Australian Research Council. *National Research Priorities and Associated Priority Goals for 2004 – 2005*. Consulted December 5 2005: http://www.arc.gov.au/grant_programs/priority_areas.htm

Fowler, Roger (1996) *Linguistic Criticism*. Oxford: OUP.

Habermas, Jurgen. *The Theory of Communicative Action*. Beacon Press, 1984.

Hirst, Martin, and Patching, Roger (2005) *Journalism Ethics: Arguments and Cases*. South Melbourne: Oxford University Press.

Kuhn, D. (1991) *The Skills of Argument*. Cambridge: Cambridge University Press.

Lamb, Matthew (2005) 'Left, Right, What's Wrong?' *Meanjin,* vol. 64, no. 3.

Niven, David (2002) *Tilt? The Search for Media Bias*.

Westport & London: Praeger, 2002.

Reed and Norman 2004, *Argumentation Machines: New Frontiers in Argumentation and*

Computation. Dordrecht: Kluwer Academic Publishers.

Reed, Chris, and Rowe, Glenn (2001) 'Araucaria Software for Puzzles in Argument Diagramming and XML.' Consulted November 15 2005: http://babbage.Computing.dundee.ac.uk/chris/publications/2001/technreport.pdf

Toulmin, S. E. (1950) *The Place of Reason in Ethics*, Cambridge: Cambridge University Press.

van Eemeren, F. H. et al (1996a) *Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments*. Mahwah, N.J.: Lawrence Erlbaum Associates.

van Eemeren, F. H. and. Grootendorst, R (1996b) 'Developments in Argumentation Theory,' in **van Benthem**, J, **van Eemeren**, F.H, **Grootendorst,** R and **Veltman,** F (eds.), *Logic and Argumentation*, North-Holland, Amsterdam, 9 – 26.

van Gelder, Tim (2003) 'Enhancing Deliberation Through Computer Supported Argument Mapping.' Consulted November 19 2005: http://www.arts.unimelb.edu.au/~tgelder/Reason/papers/Enhancing_Deliberation.pdf

Verheij, Bart (1996) *Rules, Reasons, Arguments: Formal studies of Argumentation and Defeat*. Dissertation. Universiteit of Maastricht. Consulted December 7 2005:http://www/ai.rug.nl/~verheij/publications/proefschrift/

Verheij, Bart (2001) *Legal Knowledge Based Systems/ JURIX 1999: The Fourteenth Annual Conference* (ed by **Verheij**, Bart et al), Amsterdam, Washington DC: IOS Press, 2001.

Walton, Douglas N. (1996) *Argumentation Schemes for Presumptive Reasoning*. New Jersey: Lawrence Erlbaum.

Walton, Douglas N. (1998a) *The New Dialectic: Conversational Contexts of Argument*. Toronto: University of Toronto Press.

Walton, Douglas N. (1998b) *Ad Hominem Arguments*. Tuscaloosa: University of Alabama Press.

Walton, Douglas N. (1999) *One-Sided Arguments: A Dialectical Analysis of Bias*. New York: State University of New York Press, 1999.

Walton, Douglas, and Chris Reed (2002) 'Diagramming, Argumentation Schemes and Critical Questions.' ISSA, Amsterdam. Consulted November 19 2005: http://www.computing.dundee.ac.uk/staff/creed/research/publications/2002/isse2002.pdf

Walton, Douglas, and Reed, C.A, (2005) 'Argumentation Schemes aznd Enthymemes.' *Synthese (*2005) 145: 339 – 370.

Webster, N. and Porter, N. (1913) *Webster's Revised Unabridged Dictionary of the English Language*. Springfield, MA: G. & C. Merriam.

Zukerman, Ingrid (2004) Book Review of *Argumentation: New Frontiers in Argumentation and Computation* by Reed, Chris and Norman, Timothy J. (editors), *Computational Linguistics,* vol. 31, no. 1

Digital Humanities or the *Gradiva Complex*

Séverine LETALLEUR

English Department,
Université de Paris-Sorbonne (Paris IV)

The aim of this study is to explore the mechanisms of digital contextualisation with respect to literary studies. Reversing Willard Mc Carty's definition of the field ("Humanities computing is an academic field concerned with the application of computing tools to arts and humanities data or to their use in the creation of these data"), I wish to examine in what way literary tools, in the broadest sense, apply to digital and digitised documents; and, in return what these findings add to literary studies and more specifically to literary analysis proper. More specifically, I intend to show in what respect the notion of the "Gradiva complex" in its widest acceptation (breathing life into a fixed representation such as a statue), can apply to multimedia representation through the concrete study of two different interfaces.

In reference to this, I shall first briefly comment upon the vexed link between modernity and tradition, a paradox emblematic of Humanities Computing which harmoniously blends the old and the new, giving a new lease of life to past ideas. Humanities Computing provides old representations with a brand new and dynamic frame. Such original context necessarily alters what it displays and may therefore lead to new discoveries. This is why we should take a fresh look at old methodological habits and old academic divides.

Thanks to elaborate interactive devices, what up until now seemed fixed and almost fossilised is literally brought back to life and made more accessible to each one's sensitivity. This is highlighted in the CD-ROM *Freud, Archéologie de l'Inconscient* where a sophisticated multimedia environment stages, among other things, the way in which memories are repressed. The section entitled "Gradiva" - referring to Freud's famous 1906 essay *Delusions and Dreams in Jensen's Gradiva* – digitally illustrates how Jensen's short story acts as a literary parable for the psychoanalytic process of repression, disclosing both the workings of the unconscious and those of literature - more specifically here, the metaphorical process. As we explore the CD-ROM and use the pointer to drag to the centre icons or hot-spots that would otherwise remain partly hidden in the corners[1], an eerie music starts playing. The sounds echo the whirl of smoke that runs across the screen acting as some kind of Ariadne's thread for the whole information display; it is an aesthetic and vivid reminder of Freud's presence. "Gradiva", the name given to the Pompeii bas-relief representing the protagonist's obsessive dream, is visually present at the centre but gradually buried under a pile of cinders. Meanwhile, a male voice-over reveals that the stone woman is a masked figure for a long-forgotten love. The way in which information is made accessible prevents the user from random investigation. If he or she remains passive, the screen is blurred; it is only when the user deliberately points to a specific element that the latter figure emerges from an ever-changing background (the process sometimes requires both patience and reflection, as when the next page is solely made accessible by piecing a jigsaw together). The designing choices (roll-over, invisible/visible layers) coincide with the issue at stake: the will to make sense of symptoms on the one hand, of signs and symbols on the other; a difficult quest for knowledge and understanding that mirrors both the approach of psychoanalysts and of scholars.

The digital extract briefly evoked here can be interpreted as a modern expression of the "Gradiva complex": the desire to breathe life into a fixed and cold representation. Such complex typifies the graphic designer's art or perhaps his or her fantasy. In this particular case, even though the CD-ROM is meant as a general introduction to Freud's discoveries, the fact that it relies on a multimedia support implies constant cross-references not only to different forms of representation (static or dynamic, visual or auditory), but to different sources of knowledge (cinema, literature, photography, biology, medicine) that shed light on each other and add to the revivifying effect of the tool. Digital contextualisation enhances our own understanding of each field's specific nature by hyperlinking it to what it isn't: maps, music, historical data etc. As McCarty puts it: "Humanities computing [...] is methodological in nature and interdisciplinary in scope".

In the CD-ROM *Georgian Cities*, broaching 18th century Britain, a short animation accessible in the section entitled

"A rhetorical comparison", aptly illustrates ICTs' metonymic nature: a quotation taken from Henry Fielding's *The History of Tom Jones* mentioning one of Hogarth's prints (*The Morning*) pops up unawares from each corner of the screen while the parts of the engraving successively referred to, are highlighted on the copy displayed next to the words: "[Miss Bridget] I would attempt to draw her picture, but that is done already by a more able master, Mr Hogarth himself, to whom she sat many years ago, and hath been lately exhibited by that gentleman in his print of a winter's morning, of which she was no improper emblem, and may be seen walking (for walk she doth in the print) to Covent-Garden church, with a starved foot-boy behind, carrying her prayer book." [book one, chapter XI]. The preterition device used by the narrator who initially "refers to [Miss Bridget] by "professing" not to do so, consists in the *ekphrastic* depiction of one of Hogarth's works. In the text, the witty depiction leads to the sketching-out of an almost grotesque figure. In the CD-ROM, the duplication of the print, juxtaposed with the text discloses the writer's technique. Once faced with the actual visual representation, the reader is made to understand the *ekphrasis*' literary purpose – to add distance between the subject and the imaginary object, to conceal whilst feigning to disclose. The object of literature and that of art in general remains unattainable, it lies beyond signs, beyond the visible.

In the examples aforementioned, the processes depicted seem to be duplicated by the multimedia device. The promotion of freedom that ICTs permit, along with their inherent dynamism, go hand in hand with the surfacing of cognitive processes invisible up to that point. Therein lies ICTs' main interest in terms of innovative academic research.

Humanities Computing formally overcomes the paradox between the old and the new, it also gives rise to unique questioning. It stresses the link which exists between the invisible but apt guidance involved in browsing and the interactive dimension of the tool. This link is an inextricable entanglement between apparent freedom and subtle guiding that has forever existed (notably with the literary notion of narrative viewpoint), but which is made more apparent here. Moreover, ICTs bring to light the tie that exists between creativity and learning; from objective data gathering and display, to the freest associations of ideas, thanks to interactivity. All echo the way we learn what we learn: willingly and objectively but also haphazardly and subjectively. We learn because we want to or we have to, but also because the world leaves its mark onto us.

Instead of showing how new technologies can implement humanities technically-wise, I would therefore like to show how the new medium enhances literary theory in itself; in other words - how does praxis affect theory? How from the objective compilation of documents, all rationally organised and displayed, does one move to the more irrational, and the more creative? How does one shift from the objective, the technical and the logical to the more poetical association of ideas; from knowledge acquisition to creativity? If one follows the statements put forward by the philosopher Paul Ricœur in *The Rule of Metaphor*, the metaphorical process - at the core of all literary practices - stems from the context not the actual word. It is the whole that prevails, not the part; or rather, the whole is meant in each of its parts and vice-versa. Once again, isn't it most fortunate that the new medium should now be utilized in order to enhance these very ideas?

To illustrate these points, I will use a Powerpoint presentation displaying relevant recorded extracts taken from the two cultural CD-ROMs aforementioned.

Footnotes

[1] This reversal is in itself significant.

References

Aarseth, E. (1997). *Cybertext*. Baltimore: The Johns Hopkins University Press.

Andersen, P. B. (1997). *A Theory of Computer Semiotics*. Cambridge: Cambridge University Press.

Barthes, R. (1980). *La Chambre claire*. Paris: Le Seuil Gallimard.

Barthes, R. (1977). "The photographic message". *Image Music Text*. Translated by Stephen Heath. New York: Hill and Wang.

Barthes, R. (1970). *S/Z*. Paris: Le Seuil.

Black, M. (1979). "More about Metaphor." A. Ortony. *Metaphor and Thought*. Cambridge, UK: Cambridge University Press.

Bloom, P. (1996). *Language and Space*. Cambridge, Massachusetts: MIT press.

Camara, A S. (1999). *Spatial Multimedia and Virtual Reality*. London: Taylor.

Chernaik, W., Deegan M. & A. Gibson, Eds. (1996). *Beyond the Book: Theory, Culture, and the Politics of Cyberspace*. Oxford: Office for Humanities Communication Publications 7.

Crang, M., P. Crang & J. May. (1999). *Virtual Geographies. Bodies, Space and Relations*. London: Routledge.

Cubitt, S. (1998). *Digital Aesthetics*. London: Sage.

Deegan, M. & S. Tanner. (2002). *Digital Futures: Strategies for the Information Age*. London: Library Association Publishing.

Delany, P. & G. P. Landow. (1991). *Hypermedia and Literary Studies*. Cambridge, Massachusetts: MIT Press.

Didi-Huberman, G. (2001). *L'Homme qui marchait dans la couleur*. Paris: Les Editions de Minuit.

Dodge, M. & R. Kitchin. (2001). *Mapping Cyberspace*. London: Routledge.

Fauconnier, G. (1997). *Mappings in Thought and Language*. Cambridge: Cambridge University Press.

Fielding H. (1707). *The History of Tom Jones*. (1996). Oxford: Oxford University Press.

Freud, S. (1907). *Le Délire et les rêves dans la « Gradiva » de W. Jensen*. (précédé de) Jensen, W. *Gradiva, fantaisie pompéienne*. (1986). Paris: Folio.

Foucault, M. (1966). *Les Mots et les choses. Une archéologie des sciences humaines*. Paris: Gallimard.

Gallet-Blanchard, L. & M.-M. Martinet. (2002). "Hypermedia and Urban Culture: a Presentation on the CD-Rom *Georgian Cities*". *Jahrbuch für Computerphilologie n°4*.

Gallet-Blanchard, L. & M.-M. Martinet. (2000). *Georgian Cities* cdrom. C.A.T.I. Paris: Presses de l'Université Paris-Sorbonne.

Genette, G. (1972). *Figures III*. Paris: Le Seuil.

Gunthert, A. (mai 1997). « Le complexe de Gradiva. Théorie de la photographie, deuil et résurrection». *Etudes photographiques n°2*. Paris: Société française de photographie, 115-128.

Kristeva, J. (1980). *Desire in Language. A Semiotic Approach to Literature and Art*. New York: Columbia University Press.

Landow, G.P. (1992). *Hypertext. The Convergence of Contemporary Critical Theory and Technology*. Baltimore: The Johns Hopkins University Press.

Michaud, P.- A. (1998). *Aby Warburg et l'image en mouvement*. Paris: Macula.

McCarty, W. (1998). "What is Humanities Computing? Toward a Definition of the Field."

http://www.kcl.ac.uk/humanities/cch/wlm/essays/what>

Richards, I.A. (1936). *The Philosophy of Rhetoric*. New York: Oxford University Press.

Ricœur, P. (1975). *La Métaphore vive*. Paris: Le Seuil.

Sutherland, K. & M. Deegan. (1997). *Electronic Text. Investigations in Method and Theory*. Oxford: Clarendon Press.

Sigmund Freud, Archéologie de l'Inconscient. *Portrait du Herr Doktor* cdrom. (2000). Paris: Syrinx/Le Seuil Multimédia.

Tolva, J. "Ut Pictura Hyperpoesis: Spatial Form, Visuality, and the Digital World" http://www.cs.unc.edu/~barman/HT96/P43/pictura.htm

Familles narratologiques et balisage du roman contemporain

Denise MALRIEU

CNRS-MODYCO PARIS X
dmalrieu@u-prais10.fr
Site : infolang.u-paris10.fr/modyco

The present study tries to define the novels description dimensions with a point of view that joins texts profilage and textual linguistics, with the aim to reach a linguistic characterization of novels narrative families.

The privileged description dimensions relate to the novel's enunciative device, the main actants designation, the diegetic levels and their attributes, the reported speech, their speakers and introducing segments, the focused descriptions.

The chosen methodology is the TEI/XML one, considering her interoperability. We expand the TEI tags, specially the <textClass> and <profileDesc> and define a new tags body in order to describe the previously defined categories.

We then give an example that runs with the defined tags in a contrastive study of the two types of narrator in *Le Ravissement de Lol V. Stein* of M. Duras.

At last, we underline the questions regarding the exploitation of fine tagging : the tags overlap, the readability of tagged texts and the problems generated by the mixing of different semiotic tagging layers; we emphasize the default of convivial query tools for a topologic tags analysis and for a tokens characterization by their paths in the tree.

L'exposé qui suit s'efforce de définir les dimensions de description du roman dans une perspective qui souhaite joindre profilage des textes et linguistique textuelle pour une caractérisation linguistique de familles narratologiques.

Les dimensions de description privilégiées portent sur le dispositif énonciatif du roman et concernent la désignation des actants principaux, les niveaux de diégèse et leurs attributs, les séquences de discours rapportés, leurs locuteurs et segments introducteurs, les descriptions focalisées.

La méthodologie choisie est celle de la TEI par sa proximité avec XML et les avantages d'interopérabilité. Nous avons enrichi les balises de la TEI concernant le <textClass> et le <profileDesc> et défini un nouveau corps de balises pour décrire les catégories définies plus haut.

Nous donnons ensuite un exemple d'exploitation de textes XMLisés par une analyse contrastive des deux parties correspondant à deux types de narrateur dans *Le Ravissement de Lol V. Stein* de M. Duras.

Nous soulignons enfin les problèmes liés à l'exploitation de balisages assez fins : problèmes des chevauchements de balise pour lesquels la solution n'est pas consensuelle, problème de la lisibilité des textes balisés et des désavantages d'un mélange des différentes couches sémiotiques du balisage, problème du manque d'outils conviviaux pour l'analyse topologique des balises et pour la caractérisation des occurrences par leur chemin dans l'arborescence.

L'exposé qui va suivre s'inscrit dans une démarche initiée en 2000 sur la caractérisation linguistique des genres écrits, à l'intérieur d'une linguistique textuelle et mettant en jeu la méthodologie TEI de balisage des textes. Il sera centré sur la définition des dimensions nécessaires à la description du genre romanesque et donc sur la création de balises non définies par la TEI; il fera des propositions concernant le développement de nouvelles modalités de questionnement, devenues nécessaires dès que le balisage des séquences textuelles dépasse en richesse et complexité le simple balisage des . Je terminerai par un exemple d'exploitation de roman balisé à l'intérieur du corpus des oeuvres écrites de Duras, en cours de constitution.

1 – Bref historique de la démarche

Cette démarche part du présupposé que, sur le versant littéraire, les recherches en poétique ou stylistique ne peuvent que s'appuyer sur une analyse linguistique de la matérialité des textes et que, sur le versant linguistique, l'analyse sémantique d'un énoncé implique une linguistique textuelle, qui va définir comment les traits génériques d'un texte vont contraindre l'analyse syntaxique de la phrase. Le genre apparaît comme un concept

nécessaire à cette démarche, car le genre est le lieu où s'articulent les contextes d'énonciation et les normes langagières; c'est le lieu où s'explicitent les jeux des différentes sémiotiques qui informent le texte et contraignent les dimensions cognitives de l'interprétation. L'objectif poursuivi consiste à établir un pont entre la démarche de profilage et la linguistique textuelle.

Peut-on caractériser linguistiquement des familles narratives à l'intérieur du roman?

La démarche inductive de profilage de 2540 textes à partir de 250 variables morpho-syntaxiques issues de l'analyseur Cordial[1] (Malrieu et Rastier, 2001), efficace pour différencier les domaines et champs génériques s'est avérée peu probante pour dessiner des familles narratives au sein du roman "sérieux" (le premier facteur n'expliquant pas plus de 30% de la variance).

La démarche inductive de classification des romans rencontrait plusieurs obstacles :

- non disponibilité de corpus *raisonnés* de romans contemporains suffisamment étoffés.
- nécessité de repenser les variables morphosyntaxiques à prendre en compte dans la classification (insuffisance des variables disponibles dans Cordial : nous en avons redéfini environ 300).
- le profilage permet de catégoriser mais non de comprendre les effets du texte: les calculs obtenus sur le texte dans sa globalité effectuent un lissage qui dilue les différences; le roman est par essence un texte composite (par ex le fort taux de 1PS ne peut différencier le roman homodiégétique et le roman hétérodiégétique fortement dialogué). De plus, la désambiguïsation d'un énoncé implique la prise en compte de séquences inférieures au texte, séquences informées par le genre mais qu'il faut caractériser en tant que telles.

D'où l'abandon temporaire de la démarche inductive sur textes entiers pour passer à une représentation qui prenne en compte le genre comme dispositif énonciatif.

2. – Le genre comme dispositif énonciatif :

En reformulant les hypothèses de Bakhtine, on peut avancer que le genre définit des contraintes interprétatives par un réglage du dispositif énonciatif , lié aux rapports sociaux et au dispositif communicationnel; il s'exprime dans des structures textuelles préférentielles et dans la prédilection pour certains types d'énoncés.

Dans le roman, ce dispositif énonciatif est configuré par un projet esthétique (Danon-Boileau, 1982), à la fois dans le mode d'allocution narrateur narrataire, dans la mimésis (les différents discours rapportés, leur poids, leur agencement) (Malrieu 2004), dans le mode de référenciation (dénomination vs désignation qualifiante ou anonymat) (cf plus bas une analyse contrastive des deux types de narrateur dans *le Ravissement de Lol V. Stein* de M Duras).

La nature de l'univers fictif dépend du projet esthétique de l'auteur. Celui-ci contraint la triangulation des rapports narrateur-lecteur, narrateur-personnages, lecteur-personnages. Le roman peut aménager des scènes énonciatives assez diverses, qui dépendent de la place du narrateur dans le récit, de son ethos (narrateur omniscient ou pas, distancié ou empathique, critique de façon explicite ou indirecte, menant un récit anachronique ou pas).

3 - Les dimensions de description et la méthodologie de balisage:

Il s'agit donc de définir les dimensions de description de ces scènes énonciatives qui peuvent être instables à l'intérieur d'une oeuvre. (On laissera ici de côté le dispositif particulier du roman épistolaire)[2]. La méthodologie de balisage choisie est celle de la TEI, car elle présente les avantages d'un standard international lié à XML. Le balisage de la structure arborescente du document est donc celui défini par la TEI; mais nous avons ajouté, en fonction des besoins de description, de nouvelles balises :

- *Les diégèses* : celles-ci sont décrites pour le texte entier dans le <profileDesc> et balisées dans le <body>: nombre de niveaux d'enchâssement de diégèses; nombre de diégèses de chaque niveau; longueur moyenne des diégèses de chaque niveau; chaque diégèse est balisée et décrite par ses attributs : son niveau, type de narrateur (intra- vs extradiégétique; hétéro- vs homodiégétique), l'identité du narrateur et des personnages appartenant à la diégèse, temps du récit de la diégèse, sa longueur.
- *Les focalisations* : descriptions focalisées avec identité du foyer optique.
- *Les registres de la parole* : la parole prononcée

(discours direct ou rapporté), la parole intérieure (sous-conversation consciente); leurs marquages explicites sont liés soit à la ponctuation (tiret ou guillemets), soit à des introducteurs du dire (verbes introducteurs ou en incise).

Concernant *les discours rapportés* (DR), on a défini :

- La liste des types de discours rapportés avec un who (identifiant), un speaker éventuel, un whom allocutaire explicite[3] ; les segments (le plus souvent propositions ou incises) introducteurs de discours rapportés.

(L'automatisation du balisage du DD paraît de prime abord la plus facile ; cependant il est nécessaire d'affecter préalablement le texte à la famille de marquage adéquate (familles que nous avons définies par ailleurs).

<q type="DD"> désigne le discours direct.

<q type="DI"> désigne le discours indirect.

<q type="DDR"> désigne le discours direct rapporté.

<q type="DRN"> désigne le discours rapporté narrativisé.

<q type="DIN"> désigne le discours indirect narrativisé.

<q type="MI"> désigne le monologue intérieur en DD</catDesc

<q type="MII"> désigne le monologue intérieur en DI.

<q type="MIN"> désigne le monologue intérieur narrativisé.

<q type="disc_rapp"> désigne le discussion rapportée résumée.

<q type="soCalled"> désigne le citation non prise en charge, modalisation autonyme. On éclate donc le discours indirect libre en MIN et DRN.

<ab type="TdP"><desc>description d'un *tour de parole* dans une séquence dialoguée par les attributs facultatifs </desc>:

who : désigne le locuteur du tour de parole

whom : par défaut l'interlocuteur

activity : activités et mouvements du locuteur

<seg ana="geste"> : gestuelle liée à la parole

<q type="DD" rend="tiret">contenu des paroles prononcées en discours direct</q>

Les introducteurs de DR : <seg ana="incise_di"> balise l'incise de dire; <seg ana="int"> balise les introducteurs de DD ou le DI

<seg ana="incise_pe"> : balise l'incise de discours intérieur (MI).

L'attribution dans le roman du locuteur d'un tour de parole dans les dialogues doit pouvoir être partiellement automatisé à partir d'une description des familles de marquages du DD selon les périodes et auteurs.

- *Le psycho-récit* : description par le narrateur de la vie mentale et affective du personnage sans reproduction de celle-ci; le codage du psycho-récit est effectué sur la base de présence dans la phrase de lexique lié à la cognition ou à la vie affective et émotionnelle du personnage, autre que l'expressivité gestuelle.

- La description des personnages du roman dans <particDesc> : dénomination des personnages et des locuteurs : on déclare les personnages par leur nom propre ("reg"), par le *key*, par leur rôle, on note la d'apparition et de disparition du personnage et /ou de sa dénomination par un nom propre dans le texte. L'établissement des *actants principaux d'une diégèse* se fait par repérage des noms propres humains et des noms communs avec déterminant défini singulier les plus fréquents dans la diégèse. On a ainsi les personnages principaux sans nom propre et les actants non humains principaux (qu'on peut aussi identifier par calcul des spécificités, par ex. dans *Moderato Cantabile* : la fleur de magnolia, la musique, la mer).

Un exemple de profileDesc pour *Moderato cantabile*:

```
<profileDesc>
        <textClass>
                <domaines>littérature</domaines>
                <gn>ro.ps</gn>
        </textClass>
        <totDieg nbNiv="2"/>
        <diegNiv2 nb="1" surf="?"/>
```

<diegNiv1 id="1" narr="extra_hétérodiég" narTemp="pas">

<particDesc><personae key="ad ch pa enf mg pa in"/> </particDesc> </diegNiv1>

<diegNiv2 id="2" narr="intra_hétérodiég" narTemp="pas pres"><particDesc> <personae key= "Fas Has"/>

<particDesc><personae key="pa Fas Has"/></particDesc></diegNiv2

<seqNiv2 id="2" nb="?" lm="?"/>
< catDesc>nombre de séquences de la diégèse "2" de niveau 2; longueur moyenne de ces séquences</catDesc>

</totDieg>

Un extrait de texte balisé :

<ab locus="appartement_de_mg">

<q type="DD" rend="tiret">- Veux-tu lire ce qu'il y a d'écrit au-dessus de ta partition? <seg ana="incise_di"> demanda la dame. </seg></q>

<q type="DD" rend="tiret">- Moderato cantabile<seg ana="incise_di">dit l'enfant </seg>.</q>

<seg ana="geste">La dame ponctua cette réponse d'un coup de crayon sur le clavier. </seg> <seg ana="geste">L'enfant resta immobile, la tête tournée vers la partition. </seg>

<q type="DD" rend="tiret">- Et qu'est-ce que ça veut dire, moderato cantabile?</q>

<q type="DD" rend="tiret">- Je ne sais pas.</q>

<seg ana="geste">Une femme, assise à trois mètres de là, soupira.</seg>

<q type="DD" rend="tiret">- Tu es sûr de ne pas savoir ce que ça veut dire, moderato cantabile? <seg ana="incise_di"> reprit la dame. </seg></q>

L'enfant ne répondit pas. <seg ana="geste"><PR>La dame poussa un cri d'impuissance étouffé,</PR> tout en frappant de nouveau le clavier de son crayon.</seg> <seg ana="geste">Pas un cil de l'enfant ne bougea.</seg><seg ana="geste"> La dame se retourna.</seg>

<q type="DD" rend="tiret">- Madame Desbaresdes, quelle tête vous avez là<seg ana="incise_di">, dit-elle</seg>.</q>

Balisage des diégèses dans le <body>:

<diegNiv2> <q type="DD" rend="tiret">- Ce cri était si fort que vraiment il est bien naturel que l'on cherche à savoir. J'aurais pu difficilement éviter de le faire, voyez-vous. </q> </diegNiv2>

Elle but son vin, le troisième verre.

<diegNiv2> <q type="DD" rend="tiret">- Ce que je sais, c'est qu'il lui a tiré une balle dans le cœur.</q> </diegNiv2>

Deux clients entrèrent. Ils reconnurent cette femme au comptoir, s'étonnèrent.

<diegNiv2> <q type="DD" rend="tiret">- Et, évidemment, on ne peut pas savoir pourquoi ?</q> </diegNiv2>

4 – L'exploitation du corpus XMLisé

L'exploitation des textes balisés reste malheureusement encore très peu ergonomique et conviviale pour le public littéraire ou linguistique et bon nombre de fonctionnalités souhaitables ne sont guère accessibles pour le moment.

- **On définit d'abord l'espace de recherche** : corpus entier *vs* sous ensemble de textes défini par un (ou +) trait(s) balisé(s) dans l'en-tête *vs* un texte. Ex : *l'étude des modaux dans le roman* implique de prendre en compte d'un côté (modaux) type de verbes, temps et personnes verbaux; de l'autre (corpus) de prendre en compte le champ générique et le genre (balisé dans l'en-tête dans <textClass>) et dans le roman i) le temps du récit de la diégèse en cours; ii) la séquence en cours : type de DR *vs* discours narratorial.
- **Analyses statistiques contrastives** : On peut donner quelques exemples d'exploitations pratiquées sur le roman :

- type de narrateur et surface relative des différents DR (Malrieu, 2004)
- les données statistiques morpho-syntaxiques peuvent diagnostiquer certains traits narratologiques : l'étiquetage des verbes du seul discours narratorial permet de voir s'il s'agit d'un récit au passé ou au présent; l'examen sur les différentes div de ce discours permet de voir si ce temps du récit est stable ou pas. De même le poids de la 1S dans ce discours narratorial permet de dire s'il s'agit d'un récit homodiégétique.
- Les modes de désignation des personnages: dénomination vs désignations autres selon les séquences textuelles.
- Sur un corpus important de romans, et dans une optique plus socio-historique, on peut envisager toutes sortes d'exploitation statistique de données synchroniques ou diachroniques, sur les genres, les actants, les lieux, les moments privilégiés des diégèses, etc.

- **L'analyse topologique des balises**: ce genre d'analyse n'est pas pratiqué sur l'écrit, (il l'est davantage sur les corpus oraux), il présente un grand intérêt et demanderait le développement de fonctionnalités spécifiques: les cooccurrences, positions respectives, enchaînements, rythmes, répétitions de balises fournissent des informations précieuses sur le fonctionnement du texte et peuvent donner lieu à des visualisations graphiques résumant la répartition d'un phénomène ou d'une classe de phénomènes dans le texte: ex dans le roman, les enchaînements gestuelle/psycho-récit; psycho-récit/monologue intérieur/ DD; sans parler des rythmes prosodiques.
- **Concordanciers enrichis** : le balisage des séquences textuelles permet un saut qualitatif dans la caractérisation des occurrences d'un phénomène étudié : en effet, on ne dispose plus seulement du contexte d'une occurrence dans une fenêtre de taille *n*, mais on doit pouvoir caractériser chaque occurrence par l'information sur ses contextes ascendants ou chemins dans l'arborescence: ex : occurrences de *on* sujet de telle catégorie de verbe à tel temps, selon les contextes: proposition principale ou relative ou conditionnelle; phrase interrogative vs déclarative; type de discours rapporté vs narratorial; paragraphe d'introduction de chapitre, etc; ex : densité des *on* selon les chapitres du roman et leurs types de séquences dominantes.
- **Aide à la désambiguïsation par la prise en compte du contexte balisé**: On peut considérer que la prise en compte de l'arborescence des unités supraphrastiques pour l'aide à la désambiguïsation des acceptions (ou des valeurs des temps verbaux par exemple) est un domaine encore en friche : le problème du coût de constitution des corpus finement enrichis fait pencher la balance vers les traitements statistiques lourds sur corpus pauvrement annotés. Mais il n'est pas évident que le coût final de ce dernier choix soit moindre, et la mutualisation de corpus richement annotés permettrait des exploitations illimitées dans les différentes disciplines des sciences du langage ou littéraires. ex : les modes de *résolution de la référence anaphorique* diffèrent selon que l'on est dans une séquence dialoguée ou dans le discours narratorial. ex : *valeur des temps verbaux* : si on est dans un genre narratif, si le temps du récit de la diégèse en cours = passé, si le conditionnel est dans un DI, alors le conditionnel a de fortes chances d'être un futur du passé.

5 - Un exemple d'analyse comparée des narrateurs à l'intérieur d'un œuvre de M. Duras

Les ruptures narratoriales à l'intérieur d'un œuvre ne sont pas rares chez Duras. Nous allons essayer de qualifier linguistiquement les deux configurations successives à l'intérieur du *Ravissement de Lol V. Stein*[4]: d'abord narrateur intra-hétérodiégétique, puis intra-homodiégétique. Ces deux configurations n'autorisent pas le même psycho-récit (Cohn). L'intradiégéticité met en scène un narrateur qui peut à la fois affirmer son non savoir sur le passé de Lol et impliquer le lecteur dans ses interrogations et s'adonner à un psycho-récit très empathique et projectif qui dépasse largement la parole intérieure ou le savoir du personnage sur lui-même. Le passage au récit homodiégétique affirme la thèse épistémique de Duras : il n'y a de connaissance de l'autre que dans la relation : le récit de cette deuxième partie favorise l'emprise du lecteur en combinant le mimétisme (récit au présent, dialogues en discours direct), et le psycho-récit du vécu du narrateur dans ses réactions à

Lol, les deux étant intimement mêlés.

Pour comparer les deux parties, nous avons balisé selon la méthodologie de la TEI-XML[5] les différents discours rapportés, les séquences narratoriales (intradiégétiques homo- vs hétérodiégétiques) ainsi que le psycho-récit[6].

Les variables utilisent les sorties de l'analyseur CORDIAL.

Résultats :

Poids des différents discours dans les deux parties

Par rapport à la surface totale en nombre de mots de chaque partie, le récit évènementiel (ce qui n'est ni discours rapportés (DR), ni psycho-récit (PR)) occupe des surfaces similaires (environ 40%). Le PR occupe une place plus importante dans le partie hétérodiégétique (38,5% contre 22%), le credo du narrateur sur Lol. La partie homodiégétique comporte davantage de discours direct (DD : 23% de la surface contre 2%) par contre les DR narrativisés sont moins importants (DRN 40%, MIN 15% et le DI 20% dans l'hétérodiégétique[7] n'atteignent pas 10% dans la partie Ho).

La partie homodiégétique connaît une configuration déictique cohérente : présentification du récit, dont le poids du DD est l'expression, récit au présent, marques déictiques spatiales (immédiateté du corporel et du visuel, qui embarque le lecteur dans un espace partagé).

Les temps verbaux :

Comme le montre le Tableau 1, l'homodiégétique est massivement dans le présent[8], le passé (IMP, PS et PQP) est plus important dans le récit hétérodiégétique (récit Hé).

	PR-Ho	PR-Hé
Présent indic	70,2	41,4
Imparfait	5,7	21,9
PS	1,5	11,4
PC	10,5	5,7
Futur	4,2	0,2
Conditionnel	3,4	3,6
PQP	0,5	7
Impératif	0,8	0,7

Tableau 1 : Les temps dans les deux psycho-récits

Les propriétés de la phrase : longueur et ponctuations faibles

La distinction du récit évènementiel (récit tout court) et du PR permet d'analyser en quoi l'expressivité des affects et l'idéalisation dans le PR sont corrélés avec des rythmes de phrase différents du récit et d'explorer si ces derniers diffèrent dans l'homo- et l'hétérodiégétique.

Nous avons observé *la répartition des longueurs de la phrase* : le PR, plus nettement le PR Hé, connaît des phrases plus longues que le récit, ce qui confirmerait le caractère plus élaboré du PR dans la partie Hé (plus fort % de substantifs et adjectifs/ mots signifiants, de noms abstraits, de pronoms relatifs); l'hétérodiégétique (récit comme PR) connaît des phrases plus longues que l'homodiégétique. On observe davantage de phrases très brèves dans le récit Ho (10,4% contre 2%). Les phrases ≤ 7mots représentent 18,5% dans le PR Hé contre 30,6% dans le PR Ho.

	PR-Ho	PR-Hé	Réc-Ho	Réc-Hé	Hiérarchie
Quartile 1	6	8	5	7	PR Hé > autres
Médiane	12	14	8	12	PR Hé > autres
Quartile 3	21	24	13	18	PR > Récit

Tableau 2 – Répartition des longueurs des phrases en nombre de mots

Nombre de virgules et longueur des phrases :

a) récit homo vs hétérodiégétique :

nombre de virgules : le récit Ho comporte davantage de phrases sans virgule (63,3% contre 52,8%), le récit Hé plus de phrases de 2 à 3 virgules.

b) PR homo- vs hétérodiégétique :

Nombre de virgules : le PR Ho, qui connaît des phrases plus courtes, comporte davantage de phrases sans virgules, moins de phrases à 2 ou plus de 4 virgules, sans exclure l'existence de phrases de 8 à 18 virgules. Les phrases à plus de 5 virgules représentent 4,5% dans le PR Ho contre 6,5% dans le PR Hé.

Les différences entre homo- et hétérodiégéticité

montrent entre autres que l'homodiégétique est dominé par des phrases brèves, il est moins dans la représentation au passé d'un point de vue externe, plus dans la vivacité des évènements au présent. L'homodiégéticité induit une autre forme de psycho-récit : elle autorise moins l'intrusion du discours narratorial de l'auteur et le psycho-récit s'inscrit par petites touches, par phrases brèves à l'intérieur du récit évènementiel ou des dialogues. Cependant elle n'interdit pas la phrase fortement rythmée, exprimant l'exacerbation des affects (cf. l'apparition de la virgule dans des phrases plus courtes et le plus grand nombre de virgules dans les phrases longues dans le psycho-récit homodiégétique).

Les indices thématiques : fréquence comparée des lexèmes

La comparaison des lexiques des différentes parties vient confirmer ces différences[9].

Les calculs suivants portent sur les lemmes et non sur les formes. Le % d'hapax /au nombre de lemmes différents de chaque récit ou PR montre davantage d'hapax dans les PR (60% contre 52 et 57% dans le récit homo et le récit hétéro). On peut donc dire que le psycho-récit est le lieu de la richesse lexicale, ce qui n'a rien de surprenant, vu son affinité avec la prose poétique.

Les lemmes statistiquement plus fréquents dans les PR par rapport aux récits correspondent à un lexique ayant trait aux affects (*absence, aimer, amour, cœur, craindre, douleur, dieu, larme, idée, ignorer, inconnu, infini, joie, mensonge, mentir, mort, nommer, ordre, penser, peur, rêve, souffrance, souffrir, souvenir, souhaiter, surprendre, tristesse*) mais aussi l'adjectif démonstratif singulier, certains connecteurs (*alors que, mais*).

L'examen du lexique plus fréquent dans le psycho-récit Hé / total indique un déficit pour les pronoms personnels 1S, et une surreprésentation des adjectifs démonstratifs et interrogatifs, de l'exclamation, qui marquent l'implication du narrateur, des privatifs (*ignorance, immobilité, immuable, impossible, inconnu, inconsolable, infini, infirme, inimportance, vain*), des références au monde distal (*infini, monde, univers, murer, terre, dieu, pénétrer, nom, nommer, avenir, à jamais*), des métaphores (*aurore, cendre, port, hiver, lumière, été, jour, nuit, navire, blancheur d'os, souffle, trou, mot-trou, mot-absence, gong* etc).

Les spécificités du lexique du psycho-récit Ho par rapport au lexique total (récit + PR) concernent, à côté du pronom personnel ou possessif 1S, les pronoms quantificateurs (*tout* et *rien*) ou les adverbes évaluatifs (*trop, si*), le lexique affectif (*larme, mentir, mensonge, souhaiter, joie, amour, peur, épouvante, rêver, sentiment, victoire, cœur, mort, souffrir, aimer, apaiser, accueillir, broyer, confiance, détruire, douceur, effrayer, horreur* etc), le lexique épistémique (*version, comprendre, ignorer, sens, savoir, se tromper*).

Conclusion

S'il est bien clair que le *literary computing* et plus précisément la méthodologie de balisage liée à XML et à la TEI autorisent maintenant des analyses contrastives fines des textes, il me semble que les obstacles à surmonter pour convaincre la communauté des littéraires (qui est loin de partager ce point de vue dans sa majorité) sont de quatre ordres :

- arriver à constituer des corpus balisés suffisamment volumineux et mutualiser ces ressources : pour cela il faut surmonter le goulot d'étranglement concernant le passage au format XML. Si le balisage de la structure logique du texte est évidemment d'un intérêt très limité pour l'analyse des textes littéraires, il paraît possible et nécessaire de développer des outils de balisage automatique modulables selon les genres textuels. Par exemple, le balisage de la valeur sémantique de l'italique pourrait être automatisé car son usage est réglé à l'intérieur d'un genre, de même le balisage du discours direct à l'intérieur des familles chronologiques de son marquage typographique; de même le balisage des entités nommées et des actants principaux dans le roman.
- Un travail collectif de représentation des connaissances liées à un balisage plus fin que la structure logique s'avère nécessaire; une mutualisation des représentations des balises gagnerait aussi à se définir selon les grands genres de textes (poésie, roman, articles de presse, etc).
- L'intégration des étiquetages morpho-syntaxiques et syntaxiques au sein d'un corpus XMLisé n'est pas non plus aisée pour l'instant. Le problème d'une représentation ergonomique du balisage se pose aussi et la difficulté à lire les corpus lourdement "farcis" pousserait à une séparation en portée des différentes couches de balises (structure logique, séquences

textuelles autres, balisage morpho-syntaxique, balisage prosodique etc.).

- Enfin l'exploitation des textes balisés est pour l'instant loin d'être conviviale pour le linguiste ou littéraire non programmeur, et il semble urgent de mettre à la disposition de ces communautés des outils de questionnement qui autorisent aussi bien la qualification de toute occurrence par son chemin dans l'arborescence du texte, que l'analyse topologique de la répartition des balises dans le texte, que la visualisation des occurrences d'un phénomène au sein du texte source.

Footnotes

1 http://www.synapse-fr.com

2 Il faut noter que les caractéristiques des œuvres entraînent des exigences différentes concernant le balisage : celui que je propose correspond au dispositif durassien et ne suffirait probablement pas pour décrire le dispositif de N. Sarraute.

3 Le who concerne l'identité stable et unique du personnage locuteur, le speaker concerne le rôle par lequel le locuteur est désigné (ex : Anne Desbaresdes peut être désignée par « la femme », « la mère » etc). Le whom allocutaire explicite du tour de parole ne sera dénommé que lorsque le texte le désigne, l'allocutaire par défaut étant l'interlocuteur dans la séquence dialoguée. Dans les exemples qui suivent le key des interlocuteurs n'est pas balisé.

4 Les résultats plus complets de cette étude sont accessibles sur le site : infolang.u-paris10.fr/modyco

5 http://www.tei-c.org/P4X/

6 Est balisé comme PR tout passage faisant allusion à la vie psychique, émotionnelle du personnage.

7 DRN = discours rapporté narrativisé, MIN = monologue intérieur narrativisé, DI = discours indirect, Ho = homodiégétique, Hé = hétérodiégétique; 1S = première personne du singulier.

8 Ce qui le distingue du récit homodiégétique au passé comme Le rivage des Syrtes par exemple.

9 La comparaison des fréquences des lemmes des différentes parties est faite par calcul de l'écart réduit selon la formule Z = (k-fp)/RACINE(fpq) où k = la fréquence du lemme dans la partie du texte étudiée, f = la fréquence du lemme dans le corpus de référence, p = le % total de lemmes du texte étudié par rapport au total de lemmes du corpus de référence, q = 1-p.

Références

Bakhtine, M. (1984 [1952-53]), *Esthétique de la création verbale*, Paris, Gallimard.

Benveniste, E. (1966). *Problèmes de linguistique générale*, Paris, Gallimard.

Biber, D., Conrad, S., Reppen, R. (1998), *Corpus Linguistics: Investigating Language structure and Use*, Cambridge University Press.

Bird, S. & Liberman, M. (2001). A formal framework for linguistic annotation, *Speech Communication*, 33, 23-60.

Bouquet, S. (éd.) (2004). "Les genres de la parole", *Langages, n° 153*.

Bronckart, J.P. (1996). *Activité langagière, textes et discours*, Lausanne, Delachaux et Niestlé.

Cohn, D. (1981). *La transparence intérieure*. Paris, Le Seuil.

Declerck, R. (2003). "How to manipulate tenses to express a character's point of view", *Journal of Literary Linguistics*, 85-112.

De Mattia, M., Joly, A. (2001). *De la syntaxe à la narratologie*, Paris, Ophrys.

Fillietaz, L. et Grobet, A. (1999) "L'hétérogénéité compositionnelle du discours : quelques remarques préliminaires, *Cahiers de linguistique française*, n° 21, 213-260.

Genette, G. (1972), *Figures III*, Paris, Le Seuil.

Genette, G. (1983), *Nouveau discours du récit*, Paris, Le Seuil.

Genette, G. (éd.) (1986). *Théorie des genres*, Paris, Le Seuil.

Habert, B. et coll. (2000). "Profilage de textes : cadre

de travail et expérience", *Actes des 5èmes Journées JADT*.

Hamburger, K. (1986, (trad.), [1957]). *Logique des genres littéraires*, Paris, Le Seuil.

Henrot, G. (2000), *L'usage de la forme. Essai sur* les Fruits d'or *de Nathalie Sarraute*, Biblioteca Francese, Unipress, Padova.

Kuyumcuyan, A. (2002), Diction et mention. Pour une pragmatique du discours narratif, Berne, Peter Lang.

Langue française (2001). "La Parole intérieure." n° 132.

Lintvelt, J. (1989). *Essai de typologie narrative. Le "point de vue".Théorie et analyse*. Corti. Paris.

Lips, M. (1926), *Le Style indirect libre*. Payot. Paris.

Maingueneau, D. (1986). *Eléments de linguistique pour le texte littéraire*, Paris : Bordas.

Maingueneau, D. (2000). Instances frontières et angélisme narratif, *Langue française*, 128; pp. 74-95.

Malrieu, D. & Rastier, F. (2001). "Genres et variations morpho-syntaxiques *", T.A.L., 42, 2*, 547-577.

Malrieu D. (2002). "Stylistique et Statistique textuelle: À partir de l'article de C. Muller sur les pronoms de dialogue", JADT 2002, 6èmes Journées internationales d'analyse des données textuelles, St-Malô, 13-15 mars 2002.

Malrieu, D. (2004). Linguistique de corpus, genres textuels, temps et personnes, *Langages*, *153*, 73-85.

Malrieu, D. "Discours rapportés et typologie des narrateurs dans le genre romanesque", Actes du colloque Ci-Dit de Cadiz "Dans la jungle des discours", 11-14 mars 2004, (à paraître chez l'Harmattan).

Malrieu, D. (2006). Type de narrateur et place du lecteur dans *Le ravissement de Lol V. Stein* (infolang. u-paris10.fr/modyco)

Marnette, S. (2002). "Étudier les pensées rapportées en français parlé: Mission impossible?". *Faits de Langues, 19*, p 211-20.

Marnette, S. (2002). "Aux frontières du discours rapporté". *Revue Romane*, *37.1*, p 3-30.

Marnette, S. (2001). "The French Théories de l'Énonciation and the Study of Speech and Thought Presentation". *Language and Literature*, *10.3,* p 261-80.

Mellet, S. & Vuillaume, M. (éds.) (2000). *Le style indirect libre et ses contextes*. Rodopi. Amsterdam-Atlanta.

Newman A. S. (1976). *Une poésie des discours. Essai sur les romans de Nathalie Sarraute*, Genève : Droz.

Philippe, G. (1995). "Pour une étude linguistique du discours intérieur dans "Les Chemins de la liberté" : le problème des modalités du discours rapporté", *Ritm, 11*, 119-155.

Philippe, G. (ed.) (2000). L'ancrage énonciatif des récits de fiction, *Langue française*, 128.

Rabatel, A., 1998, *La construction textuelle du point de vue*, Lausanne, Paris, Delachaux & Niestlé

Rastier, F. (2001). *Arts et sciences du texte*. Paris, PUF.

Rivara, R. (2000). *La langue du récit*, Paris, L'Harmattan.

Rosier, L. (1999). *Le discours rapporté. Histoire, théories, pratiques*. Paris, Bruxelles, Duculot

TEI, http://www.tei-c.org

TEI P5, http://www.tei-c.org/P5/

De-Constructing the e-Learning Pipeline

Jan Christoph MEISTER
Birte LÖNNEKER

University of Hamburg

For the 'hard-core' computing humanist, e-learning seems to be a non-topic.1 While foundational issues and Humanities Computing (HC) curriculum development have been central to our debates, the technological and didactical nitty-gritty of e-learning appears to offer little insight into questions of theoretical and conceptual relevance. - This contribution argues to the contrary:

(1) The current form of e-learning is shaped to a significant degree by business and political interests.

(2) Consequently, some of the dominant commercial e-learning platforms are conceptually flawed: they are implicitly based on a simplistic 'cognitive pipeline'-model of learning.

(3) HC can help to deconstruct simplistic e-models of learning and contribute towards a more 'intelligent' computational modelling of learning processes.

(4) The latter will be demonstrated by way of a practical example, the intermedial narratological e-course 'NarrNetz'.

(1) Vested interests

The impression of e-learning as a non-inspiring 'coal-face' activity is partly a result of its swift appropriation by commercial software vendors. Indeed: there's considerable money in e-learning – but where does it come from? In Europe, the EU and its member states have created various funding structures reserved for e-learning projects, notably *eLearning: a programme for the effective integration of Information and Communication Technologies (ICT) in education and training systems in Europe* which aims at the strategic proliferation of e-learning across the European school and university systems. [2]

Meanwhile, the initial euphoria about e-learning – fuelled in no small way by university administrators' and politicians' expectation to introduce a cheaper, less staff-intensive way of teaching – has certainly worn off. The new buzzword is 'blended learning': the combination of physical classroom teaching with auxiliary e-learning. According to a survey presented in Schulmeister (2005), the vast majority of e-courses at non-virtual colleges and universities is indeed of an auxiliary (preparatory, supplementary or remedial) nature. Schulmeister concludes that in our present situation "E-learning ist die Reparatur am System, das wir einführen" – e-learning is the quick-fix for the BA/MA-system whose current introduction at German universities leads to a dramatic over stretching of teaching resources.

(2) The pipeline model of learning

However, not all short falls can be attributed to context. In our own practical experience, part of the methodological and didactical problems associated with e-learning stem directly from restrictions imposed by its platforms. One of the key functionalities of a good e-learning platform is the personalized, dynamic assignment of navigation options (so-called 'content release') on the basis of results achieved by a user in preceding tests. Depending on the scoring, hyperlinks may be made visible or activated, or the system may send some additional feed-back. If the content manager decides to use these functionalities, the course content must allow for tests and parallel navigation paths. Under these conditions, this kind of interactivity can be hard-wired into the course by the content developer who uses system provided templates for content and test modules (questionnaires, multiple-choice etc.) and combines these with logical or quantitative operators that specify navigation rules.

To understand the conceptual premises built into commercial e-learning platforms (also known as LMS = Learning Management Systems), it is important to take a closer look at the kind of binary logic that drives this type of interactivity. For example, the currently still widely used WebCT Campus Edition 4 allows neither for the specification of complex (conjunctive or additive) rules of the type 'if student X passed test 1 with > 65% and has already looked at modules 15-17, allow him/her to jump to module 22, else do Y', nor for a logical 'or' in the

control of a user's navigation - fairly trivial process control decisions which a human teacher will make on the fly in any given class-room situation.[3] The impact of such concrete technological restrictions on e-learning's current shape is severe, considering that the recently merged WebCT and Blackboard alone account for a joint 2/3 market share in institutional and commercial e-learning.[4]

Courses implemented on the dominant proprietary platforms, unless their content modules are made available in pure (non-conditional) hyperlinked form, are thus implicitly based on what we call a 'pipeline' model of learning: the learning process within the system is organised in terms of a *linear accumulation of pre-defined knowledge and skills*. While the genuine dynamics of interactive learning in a real class room is the result of engagement in recursion, exploration and the adoption of a different perspective onto the same problem, systems based e-learning thus reduces much of this functionality to mere repetition ('repeat until you pass'), while outsourcing true interactivity to humans (via chats, discussion list, e-mail etc.).

Another principled problem with out-of-the-box e-learning systems is that they are themselves not designed to learn. Whereas commercial data-base driven websites like Amazon's combine continuously updated user-profiling (items bought, pages looked at, duration of visit, etc.) with probability based feed-back ('other users who have bought item X also bought item Y'), similarly intelligent behaviour has as yet not been implemented in the leading commercial e-learning or course-ware systems. While most e-learning platforms are of course able to track user behavior, they remain unable to act on process-relevant conclusions that might be drawn from this data.[5]

(3) Towards adaptive e-learning

So, is (e-)learning by necessity a dumber activity than shopping? Certainly not. Current AI research explores various possibilities of making it 'feel' more intelligent. For example, Nakano, Koyana and Inuzuka (2003) explore the role of probabilistic algorithms in the automated generation of questions; Steinemann et al (2005) as well as Yang et al (2005) look into the design principles of more sophisticated course management systems. Providing the user with intelligent corrective feed-back is the main goal of INCOM,[6] a project dedicated to developing an e-learning supported logic programming course (Le 2005, Le and Menzel 2005).

On a more general level, Moreno et al. (2005) emphasize the need for so-called 'adaptive e-learning' systems: ' *Adaptive e-learning* is a teaching system which adapts the selection and the presentation of contents to the individual learner and his learning status, his needs, his learning style, his previous knowledge and preferences.' This topic receives particular attention in the AH (Adaptive Hypermedia and Adaptive Web-Based System) conference series.[7] ALFANET, a project supported by the EU-financed IST- (Information Society Technologies) initiative, also aimed to introduce adaptivity.[8] According to the project's self-description, the prototype can dynamically evaluate a user's knowledge and then selectively assign different content material in order to reinforce or broaden a subject.[9] However, the final project report proves this to be an aim, rather than a demonstrable accomplishment.[10]

(4) Opening the black box

While we stake no claim to developing an intelligent e-learning *system*, we do believe that an individual e-learning *course* can certainly be turned into a more stimulating and challenging experience through comparatively simple system adjustments – provided the system itself is not a proprietary black box. Whether such adaptations and extensions can later be implemented as generic system functionalities remains to be seen.

'NarrNetz' – short for 'Narratology Network' – uses the widely adaptable PHP/MySQL-based open source platform ILIAS[11]. Our project is conceptually based on a learning metaphor and an architecture that combines linear modules with explorative, highly personalised learning experiences. Elements of NarrNetz try to emulate the user experience of an interactive computer game – a paradigm widely expected to revolutionize the 'look and feel' of e-learning (Begg, Dewhurst, Mecleod 2005; Foreman/Aldrich 2005; Schaffer 2005). In NarrNetz, the user is initially confronted with a problem modelled on a gaming 'quest'. However, the system and its architecture allow the combination of linear with non-linear, systematic (deductive) with immersive (inductive) learning methods. If learners want to

fast-track their progress, they can switch to a text-book like learning style. Conversely, while sections of the 'learnflow' follow a compulsory linear trajectory, others may be chosen and arranged at will. *Fig. 1* sketches out the NarrNetz-learnflow which maps the user's navigation through the landscape of the metaphorical quest onto the didactically desired cognitive progression.

In our presentation we will demonstrate how comparatively simple modifications and modular extensions of the open-source e-learning platform ILIAS and the integration of multi-media objects can lead to a significantly enriched, more 'intelligent' learning experience. Specifically, we will discuss the following additional functionalities:

- integration of user-specific avatars with dynamically attributed 'real-time' properties symbolizing the user's past progress and future navigational options at any given point in time;
- automated, personalized navigation control based on disjunctive reasoning and the concept of multiple, functionally equivalent learning paths rather than a pipeline-approach;
- representation of the newly acquired skills and knowledge elements in the form of a personalized dynamic conceptual glossary; glossary entries can be cross-referenced, individually annotated and linked to content elements of the course.

Conclusion

The conceptualisation and implementation of non-

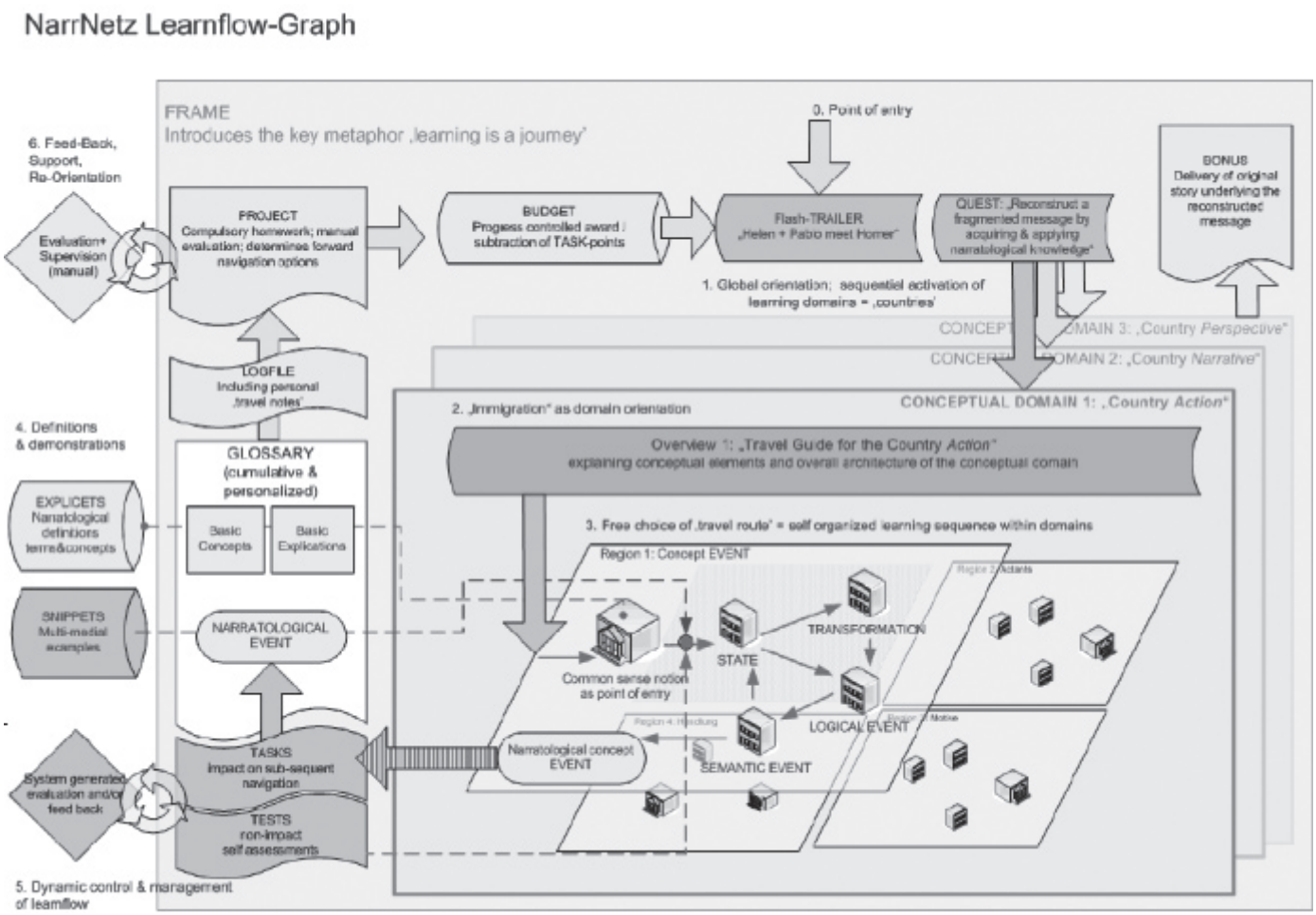


Fig.1: NarrNetz Learnflow

standard, experimental e-learning architectures constitutes a genuine task of Humanities Computing. As in other fields of applied humanities computing, computing humanists are called upon to point out the conceptual shortcomings in simplistic attempts at reverse engineering human cognitive competences for the mere sake of making them 'computable'. Learning is the most basic cognitive competence in human beings: a good enough reason not to turn a blind eye on how it is being conceptualised – technologically and ideologically.

Footnotes

1) HUMANIST vol. 1-19 contains 46 occurrences of the keyword 'e-learning' in calls for papers, conference and publication announcements, but no genuine discussions of the topic. JLLC vol. 1-17 contains 24 rather loosely related articles, three of which are of a methodological or critical nature: Koch 1992, Zuern 1994, McGrady 1999. The relevant HUMANIST (vol. 1-18) and JLLC entries are listed in www.jcmeister.de/texts/allc-2006/HUM-1-18-JLLC-1-17.doc. – For a recent non-HC affiliated evaluation of current e-learning methodologies see Meister et.al. 2004 (author not related).

2) On the EU e-learning initiative see http://europa.eu.int/comm/education/programmes/elearning/index_en.html.

3) WebCT Campus Edition 6 claims to support the combination of multiple types of release criteria, using Boolean logic (see http://www.webct.com/ce6/viewpage?name=ce6_content_presentation). However, of four classes of criteria (*date, member, group, grade book*) only 'grade book' (i.e. test results) is a dynamically instantiated variable; all others call up absolutely, manually defined attributes. More importantly, characteristics of a learner's individual navigation track cannot be specified as criteria: they are only available for manual inspection or statistical behav ioral analysis. For details see: Designer and Instructor Reference. WebCT^{TM} Campus Edition 6.0. Technical Communications Department. July 16, 2005. Chapter 33: Selective Release, 773-791.

4) See http://www.insidehighered.com/news/2005/10/13/merger

5) A word of caution: in teaching and learning, probabilistic reasoning based system's feed-back and navigation control can easily become disfunctional or meet with learners resistance. What is at stake is the very individuality of a learning experience that may very well defy statistical regularity.

6) On INCOM, see http://nats-www.informatik.uni-hamburg.de/view/INCOM/WebHome

7) See http://www.ah2006.org/. Also see the forthcoming special issue on Adaptive Hypermedia, Journal of Digital Information: http://jodi.tamu.edu/calls/adaptive_hypermedia.html

8) On IST, see http://istresults.cordis.lu/

9) See http://istresults.cordis.lu/index.cfm/section/news/Tpl/article/BrowsingType/Short%20Feature/ID/77772

10) See Barera et.al.., 2005: D66 Evaluation results, 18-21. http://rtd.softwareag.es/alfanet/PublicDocs/ALFANET_D66_v1.zip

11) See http://www.ilias.de/ios/index-e.html

References

(All weblinks accessed 08.11.2005)

Begg, M.; Dewhurst, D.; Mcleod, H. (2005). Game-Informed Learning: Applying Computer Game Processes to Higher Education. *Innovative. Journal of Online Education* (August/September 2005: http://www.innovateonline.info/index.php?view=article&id=176)

Foreman, J., and Aldrich, C. (2005). The design of advanced learning engines: An interview with Clark Aldrich. *Innovate* 1 (6).

Koch, C. (1992). Combining Connectionist and Hypertext Techniques in the Study of Texts: A HyperNet Approach to Literary Scholarship. *LLC* 1992, 7: 209-217; doi:10.1093/llc/7.4.209

Le, N.-T.; Menzel, W. (2005). *Constraint-based Error Diagnosis in Logic Programming*. 13th International Conference on Computers in Education 2005.

Le, N.-T. (2005): *Evaluation of a Constraint-based Error Diagnosis System for Logic Programming*. 13th International Conference on Computers in Education 2005.

McGrady, D. (1999). Making 'wreaders' our of students: some strategies for using technology to teach the humanities. *LLC* (1999), 14: 237-256; doi:10.1093/

llc/14.2.237

Meister, D.M. et al (eds.) (2004). Evaluation von E-Learning. Zielrichtungen, methodologische Aspekte, Zukunftsperspektiven. Münster (Waxmann) 2004.

Moreno et.al. (2005). *Using Bayesian Networks in the Global Adaptive E-learning* Process. Paper read at the EUNIS conference, June 21-25, University of Manchester http://www.mc.manchester.ac.uk/eunis2005/medialibrary/papers/paper_130.pdf

Nakano, T. (2003). An approach of removing errors from generated answers for E-learning. *17th Annual Conference of the Japanese Society for AI*. http://www-kasm.nii.ac.jp/jsai2003/programs/PDF/000083.PDF

Schulmeister, R. (2005). *Studieren in der Informationsgesellschaft. Vernetzt, modular, international?* (Presentation at the Campus Innovation 2005 conference, Hamburg, 20-21 September 2005.) http://www.campus-innovation.de/upload/dateien/texte/schulmeister_ci_2005.pdf

Shaffer, D. (2005). Epistemic games. *Innovate* 1 (6). http://www.innovateonline.info/index.php?view=article&id=79

Steinemann, M.A. et al (2005). Report on the 'Virtual Internet and Telecommunications Laboratory (VITELS)' project. *IAM Annual Report Academic Year 2004/2005*. http://www.campus-innovation.de/upload/dateien/texte/schulmeister_ci_2005.pdf

Yang, F. et al (2005). A Novel Resource Recommendation System Based on Connecting to Similar E-Learners. In: *Advances in Web-Based Learning – ICWL 2005*. Berlin, Heidelberg: 122-130.

Zuern, J. (1999). The sense of a link: hypermedia, hermeneutics, and the teaching of critical methodologies. LLC 1999, 14: 43-54; doi:10.1093/llc/14.1.43

Authors' e-mail addresses:

Jan Christoph MEISTER – mail@jcmeister.de

Birte LÖNNEKER – birte.loenneker@uni-hamburg.de

Pcube – Policarpo Petrocchi Project : The Architecture of a (Semantic) Digital Archive.

Federico MESCHINI

Tuscia University

The aim of the Pcube (Policarpo Petrocchi Project) is to create a digital archive about the cultural and intellectual production of Policarpo Petrocchi. This archive will preserve and disseminate digital objects and their associated metadata, using current technology and metadata standards and will implement features and functionality associated with the Semantic Web. Policarpo Petrocchi was an important intellectual and political figure during the second half of the ninetheenth. Petrocchi is famous as the author of the "Il Novo dizionario universale della lingua italiana" (The new universal dictionary of Italian language), used widely in the early twentieth century, and he was also the author of many literary works, reviews and translations, as the L'Assommuàr of Émile Zola, and the founder of the society "Italia e Lavoro". Petrocchi's cultural production has left us with an "information legacy" composed of extremely heterogeneous materials, which can be divided in three main categories. The first category is archival materials, including letters, press clippings and manuscripts. Second is the published works, including literary works, educational and grammar books, novels and translations. The final category is the iconographic material, composed of loose photographs and a family photo album. Moreover none of these materials has been digitized until now, and the relative metadata have been catalogued in two separated databases: the first, for the archival description, with WinISIS <http://www.unesco.org/webworld/isis/isis.htm>, using a structure that has been designed to be ISAD(G) compatible; the second, for the bibliographic records, has been created with EasyCat <http://www.biblionauta.it/biblionauta/easycat.php>, using the ISBD(G) <http://www.ifla.org/VII/s13/pubs/isbdg.htm> format, with the possibility of an UNIMARC <http://www.ifla.org/VI/3/p1996-1/sec-uni.htm> export. Even if widely used in library and archival contexts, these two databases are far from optimal starting points,

because, even if both of them can be implemented as an OPAC using plug-in and extensions, they cannot be directly integrated, and this is contrary to the goal of having a homogeneous and open system as the infrastructure for this digital archive, which must integrate electronic texts, digital images and metadata about both the analog and digital objects. Currently, the best model for this kind of infrastructure is certainly the Open Archival Information System (OAIS), with its concept of the different types of "Information Packages" and in particular the Archival Information Package (AIP). With the OAIS Model in mind the architecture of the Policarpo Petrocchi Digital Archive has been designed with three different levels. The first level is the data layer, perhaps the most delicate one, because it's where actually lie the roots for any possibility of interoperability between the different data sets. Another classification of the different materials can be made using the nature of the digital representation that will be obtained from them, being either visual or textual. All the photos, most of the letters and manuscripts, and the most important part of the literary works will be digitized in an image format, following the guidelines of the Digital Library Federation<http://www.diglib.org/>, with a high quality format for preservation and a lower quality format for dissemination. XML will be used for the digitizing the content of some of the archival documents and novels, and for the encoding of the metadata of all the analog items. EAD (Encoded Archival Description) <http://www.loc.gov/ead> is the most suitable metadata format for the archival documents, and EAD allows links to digital representations, wheter textual or iconographic, from the EAD finding aid. For published articles and books, the MODS <http://www.loc.gov/standards/mods> schema a rich bibliographic metadata format, will be used. Another issue which will be described, is the possibility of conversion into EAD and MODS from the legacy data of the databases, in a direct way, using XML export features, with some string manipulation through a programming language, or, when needed, in a manual way. The content of the books and the manuscripts will be encoded using the TEI Guidelines. The coherence and cohesion of all these standards, and of all the related digital objects, can be achieved using the METS <http://www.loc.gov/standards/mets> schema, the primary function of which is the encoding of descriptive, administrative and structural metadata of the items constituting a digital object(14). The second level is the framework layer, the implementation of the software architecture which has to provide the basic functions of a digital library, using as a base the data of the first level. During the last couple of years the number of this kind of software programs, and their availability in open-source mode, has constantly increased(15). Notwithstanding this, due to the characteristics of the Policarpo Pretrocchi Digital Archive, what is needed is an high level of customization, and the integrated use of the Apache Cocoon Framework <http://cocoon.apache.org> with the XML Database eXist <http://exist.sourceforge.net> its probably the most suitable choiche. Using these two programs, the potentialities of the XML technologies XQuery and XSLT offer a lot of possibilities, from textual and metadata researches to electronic edition with multiple output format and text/image visualization. The final level is the semantic layer, which aims to create a network of relationships among the items contained in the archive and possibly external resources and thereby to offer advanced navigation functionalities to the archive users. The ontology takes as a model the CIDOC-Conceptual Reference Model (CRM) <http://cidoc.ics.forth.gr>, which "provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation". This abstract model must be rendered to an actual syntax and for this role has been chosen the ISO standard of the TopicMaps <http://www.topicmaps.org> with the XTM serialization. The reason of this choice is the growing adoption of TopicMaps in the Digital Humanities community, compared for example to RDF/OWL. Starting from the particulare case of the Pcube project, this paper will analyze, extract and outline the general guidelines and the best strategies in the creation of a digital archive composed by very different starting materials, in order to make these strategies applicable to several other projects, which are currently facing the same issue in crossing the borders from the old models of the cultural heritage preservation towards the new paradigms of the digital library.

References

Aa. Vv. (2000). *ISAD (G) : general international standard archival description.* Ottawa, Canada.

International Council on Archives.

Aa. Vv. (2002). *Reference Model for an Open Archival Information System (OAIS)*. Wasghinton, USA. Consultative Committee forS pace Data Systems.

McDonough, J. (2002). *Encoding Digital Objects with METS*. in **Tennant, R (ed)**. *XML in Libraries*. Neal-Shuman Publisher. pp. 167-180.

Meschini, F. (2006). *TMS – TEI Management Systems*. in **Aa.Vv.** *ODOK '05*. VÖB.

Sperberg-McQueen, C. M. and Burnard, L. (eds). (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium.

Tuohy, C. (2005) *New Zealand Electronic Text Centre: Using XML Topic Map to present TEI*. TEI Members Meeting 2005 Presentation. Bulgaria.

Walsh, J. (2005).*TM4DH (Topic Maps for Digital Humanities): Examples and an Open Source Toolkit*. in *ALLC/ACH 2005 Proceedings*. Victoria. Canada.

Tagging Categorial Fuzziness and Polyfunctionality

Anneli MEURMAN-SOLIN

Helsinki Collegium for Advanced Atudies, University of Helsinki

While electronic databases have constantly improved as regards their quantitative and qualitative validity and relevance, compromises have sometimes been made in their tagging by relying on pre-corpus-linguistic grammatical descriptions, resorting to automatic (i.e., non-interactive) tagging, or imposing neat category labels on the data. One of the motivating factors in my present project, which aims at producing a web-based tagged corpus of Scottish letters (*Corpus of Scottish Correspondence*, 1500-1730, CSC), is my conviction that, besides ensuring the authenticity of data by digitising original manuscripts rather than editions, it is also necessary to create text annotation principles and methods appropriate for dealing with the great degree of linguistic variation and variability recorded in historical documents. Tagging may distort evidence by applying overly rigid rules in categorisation, so ignoring the inherent fuzziness of categories, and by simplifying complex patterns of variation or using tags that fail to reflect processes of change over a long time-span. These problems are particularly challenging in the tagging of corpora for the study of those regional and local varieties of English which may significantly differ from previously described standardised varieties. One of these is Scots, an internally quite heterogeneous variety.

In my paper I will discuss variationist principles of tagging, focusing on how multiple category membership resulting from variation over time and space has been dealt with in the system applied to the CSC. The general practice is that the tag, consisting of a lexel and a grammel, provides descriptive information about structural and contextual properties, and may contain comments permitting, for example, semantic specification or disambiguation, but the properties listed in the tag do not suggest a straightforward categorisation. Instead, the tag indicates the variational pattern that the tagged item is a member of, allowing the creation of valid inventories for

the study of the continual variation and change inherent in any variety, idiolectal, local, regional, or supraregional.

In creating variationist typologies, the deciding factor is that variants in a particular typology have been observed to show patterning at a particular level of analysis, be it structural or syntactic, or related to communicative or text-structuring functions. Thus the textual, discoursal, sentential and clausal levels are distinguished from one another. For example, subordinators such as *since*, *as*, *because* and *for* (the first two in the role of cause) and connective adverbs such as *therefore* may be members of the same semantically defined inventory, but their membership of a syntactically defined group is not as straightforward as the traditional categorisation into conjunctions and adverbs would suggest. *For* in particular requires a careful analysis of the data. When the qualifications of these items for membership are assessed at the discoursal and textual levels, there is evidence that they are not members of the same patterns of variation (Meurman-Solin 2004a).

Secondly, the system has been tailored to meet the challenge of tracing developments over a long time-span. This requires that the source of a particular item or collocate is kept transparent over time, even though a later grammaticalisation or reanalysis, for instance, would permit recategorisation. Thus, *according to* is tagged 'accord/vpsp-pr>pr and 'to/pr<vpsp-pr, *seeing that* 'see/vpsp-cj{c} and 'that/cj<, and *exceedingly* 'exceed/vpsp-aj-av and '-ly/xs-vpsp-aj-av, indicating that the ultimate source of all these is a present participle. Tags providing information about potential rather than established membership of categorial space are intended to ensure that comprehensive inventories can be created for examining the full scale of variation. For example, the tag for the variational pattern of *upon* (*the*) *condition that* includes *upon this condition that* ('condition/n-cj<pr), even though the determinative element *this* positions the variant at the end of the cline where an abstract noun followed by a nominal *that*-clause as the second unit would be analysed as an appositive structure. The approach to lexicalisation is the same. Consequently, in an invariable collocate such as *anyway* the two units are tagged separately, and their interrelatedness is indicated by arrows: 'any/aj>n-av and 'way/n-av<aj.

The third factor stressed in the theoretical approach is the inherent fuzziness and polyfunctionality recorded in language use when examined by drawing on representative large-scale corpora. The tagging principles have been influenced by the discussion of notional or conceptual properties, elaborated tags making the interrelatedness of the members of a particular notional category explicit (Anderson 1997). Even though conventional part-of-speech labels are used in the tags, strings of such labels indicate fuzziness and polyfunctionality by referring to the co-ordinates of items on a cline, rather than insisting on membership of a single category. Concepts such as 'nouniness' and 'adverbhood' reflect this approach. *Right honorable* as a term of address is tagged 'right/av, 'honour/aj-n{ho}-voc and '-able/xs-aj-n{ho}-voc ({ho} commenting on the honorific function, 'voc' being an abbreviation of vocative). The tag type with two core category names combined by a hyphen (here aj-n) permits searches concerned with fuzziness or polyfunctionality. This practice can be illustrated by *conform to*, 'conform/aj-pr>pr and 'to/pr<aj-pr, and *as soon as*, 'as/av>cj 'soon/av-cj 'as/cj<av. Ambiguity is made explicit by using comments such as {syntactic ambiguity>} and {syntactic merger>}, which alert the user to re-examine the problematic structure that follows the comment.

In applying the variationist approach to the reconstruction of historical language varieties, it is necessary to keep in mind that due to scarcity of texts, in the early periods in particular, there may be significant gaps in our knowledge base. Ideally, members of a particular pattern of variation have all been attested in the data as genuine alternatives, and can be related to one another by addresses provided by the tags. In practice, if only fragments are available, rather than a balanced and representative collection of texts, we will also have to deal with potential, though unattested, variants. In these cases, we must resort to typological information on corresponding systems in later periods, or those representing other languages or language varieties.

Therefore, another essential ingredient in the theoretical approach is that zero realisations are included in variationist paradigms. There is a wide range of zero realisation types, and the tagging principles vary accordingly. Elliptical items, including features carefully discussed in grammars such as *that*-deletion, are indicated by comments such as {zero v}, {zero aux}, {zero that}, and {zero S}, while in the relative system both a comment and a tag introduced by a zero are used: {zero rel} followed by '/0RO{y1}, for instance. Since the explicit marking of semantic relations

at various levels is a salient feature in early letters, zero marking being much less frequent in the historical data examined than in similar registers in Present-day English (Meurman-Solin 2004b), comments are also used to indicate the absence of logical connectors. It should be noted that, in principle, the term zero-realisation can be considered misleading, since it may not always be possible to verify empirically that something has been omitted. However, in the present approach a zero realisation is indicated when a variational pattern comprising explicitly expressed variants and those left implicit has been repeatedly attested in the data.

References

Anderson, J. M. (1997). *A Notional Theory of Syntactic Categories*. Cambridge: Cambridge University Press.

Jackendoff, Ray. 2002. *Foundations of language: brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Meurman-Solin A. (2004a). Towards a Variationist Typology of Clausal Connectives. Methodological Considerations Based on the Corpus of Scottish Correspondence. In Dossena, M. and Lass, R. (eds), *Methods and Data in English Historical Dialectology*. Bern: Peter Lang, pp. 171-197.

Meurman-Solin A. (2004b). From inventory to typology in English historical dialectology. In Kay, C. J., Horobin, S. and Smith, J. J. (eds), *New Perspectives on English Historical Linguistics*, Vol. I: *Syntax and Morphology*. Amsterdam / Philadelphia: John Benjamins, pp. 125-151.

Writing With Different Pictures : New Genres for New Knowledges

Adrian MILES

Applied Communication. RMIT University

Note: This is an abstract only and it refers to a more complex audiovisual essay that will form the heart of this 'paper'. A early version of such a work (which forms part of a different networked videographic essay) is available to be viewed if necessary. It is an archive (http://hypertext.rmit.edu.au/downloads/ACH06.zip) which requires QuickTime 7.x or better, to view and is 10MB in total size. I provide this to give some indication of the sort of work that will be written and presented for this paper.

The SMIL Annotation Film Engine (presented at Digital Resources in the Humanities 2002) was a middle scale computing humanities project that sought to combine a real time video stream, an idiosyncratic metadata scheme, and the annotation and presentation affordances of SMIL.

The first version of this project "Searching", was based on John Ford's 1956 western _The Searchers_ and a metadata scheme was defined from a hermeneutic claim, that "doorways in _The Searchers_ represent liminal zones between spaces that are qualities". Doors, as they appear in the film, were encoded around a small data set (camera is inside, outside, or between, and is looking inside, outside, or between) and still images from the film are provided. A search by a user yields all the stills that meet the search criteria, and clicking on any still loads the appropriate sequence from the film for viewing in its cinematic context.

As I argued in 2002, "Searching" operated as a 'discovery engine' rather than a metadata archival project, so what it developed in the engine is a process for the unveiling of patterns of meaning, and this would seem to offer an engaged middle ground for distributed humanities content that provides access to material while also foregrounding an applied critical or interpretive activity.

Since 2002 such idiosyncratic data schemes have emerged via social software as folksonomies, and I have adopted the implications of this earlier project to develop a series of interactive video 'sketches' that foreground and explore the implications of a critical videographic essay practice. This is the writing of video, sound, text, and image into audiovisual knowledge objects that are porous to the contemporary network. These works have implications for new media and humanities pedagogy, as they offer a possible heuristic for teaching and learning that is between the traditional essay and the 'interactive' (whether online or CDROM hardly matters) reference work.

These sketches are academic works that aim to reverse the usual hierarchy (and ideology) of relations between text and image that exists in academic research by allowing video and sound to 'drive' the writing (and vice versa). The model proposed allows academics or students to write their own interactive video based essays that are multilinear, academically sound, and network appropriate.

The work has implications for humanities disciplines that study time and image based media, new media theory, and contemporary multimedia pedagogies, and the 'paper' will be such an 'networked knowledge object'.

Digital Gazetteers and Temporal Directories for Digital Atlases

Ruth MOSTERN

Founding Faculty
School of Social Sciences, Humanities and Arts University of California, Merced

Digital gazetteers—databases of named places--are becoming widely recognized as an integral element of any application or library that includes geographical information. In the humanities, where historical maps and texts are replete with place names but precise spatial information is often limited, it is often a good practice to conceive of the development of geographical information as beginning from place names (a gazetteer) rather than geometry (a GIS).

As a reference work, digital gazetteers are an exceptionally useful tool in their own right, making it possible to trace the history of named places, relate places to one another, link multiple names for the same place across languages, and (with even rudimentary georeferencing), and make maps. As a component of a digital library or web application, gazetteers permit place-based or map-based search, display and integration of any kind of content that includes named places. Finally, using existing service protocols, multiple gazetteers can be searched and used together, allowing very specialized research in historical geography to be used in new contexts.

This paper has several components. First, I introduce recent research in digital gazetteers for digital libraries and the humanities, with a primary focus on work by the Alexandria Digital Library and modifications to its standard by the Electronic Cultural Atlas Initiative.

Second, I work through several examples of gazetteer design from my personal research in early modern Chinese history, looking at frequently changing places in a time of rapid political change, complex and localized religious spatial hierarchies, and multiple, coexisting civil and military spatial systems. I reflect on a classic cultural geography notion of space from Yi-fu Tuan

that "as centers of meaning, the number of places is enormous, and cannot be contained in the largest gazetteer." In response to Tuan, who made this claim in 1975, I respond both that more recent developments in database design and networked systems mitigate the "largest gazetteer" concern; but also, I claim that it is possible—and, for use in the humanities and related fields of cultural heritage management, essential—to create a gazetteer that acknowledges places as centers of meaning, not just administrative spaces.

Finally, I discuss on emerging work by myself and other colleagues in the design of named time period directories. While digital gazetteers are a well-recognized genre (indeed, their antecedents begin long before the digital era), databases of time periods are not. Nevertheless, search, display and integration of historical and cultural information demands that such a resource be developed. Times, like places, have names—overlapping, multiple, hierarchical, and rich with information. Just as time modifies place (as cities are founded, rivers change course, or empires are vanquished), so too does place modify time (as the Abbasid Caliphate refers to a political formation on a particular part of the earth at a given time, or the Neolithic Period begins and ends at different times in, say, the Yellow River valley or the Andes). This paper introduces some of the issues concerned with data modeling and visualization of time period directories along with gazetteers in digital atlases, and discusses next steps.

L'apport pédagogique de la vidéo dans le module « Chairing a meeting »

Marc NUSSBAUMER

Université Nancy 2, France.

C'est en réponse à un appel d'offre de LUNO (Lorraine Université Ouverte) que l'équipe enseignante du CTU (Centre de Télé-enseignement Universitaire) de Nancy 2 a proposé de créer le module « Chairing a meeting in English / Conduire une réunion en anglais ».

LUNO est un programme de formation ouverte et à distance conduit par les établissements d'enseignement supérieur lorrains regroupés autour d'une charte et initié par le Conseil Régional de Lorraine en septembre 2000. Ces établissements ont développé dans leurs champs de compétence respectifs, des modules de formation utilisables à distance et accessibles à tout demandeur d'emploi ou salarié lorrain.

Le module « Chairing a meeting in English » est destiné à des étudiants en formation continue pouvant suivre une formation en ligne. La spécificité du public fut le premier critère à prendre en compte dans l'avant-projet. Contrairement aux étudiants « classiques » dont la motivation est extrinsèque (objectif premier : obtention du diplôme), ces apprenants ont une motivation plus intrinsèque (objectifs : plaisir d'apprendre, satisfaction personnelle) qui est considérée comme le plus haut niveau de motivation autodéterminée. Nous souhaitions toutefois susciter davantage encore l'intérêt de ce public en adaptant les outils aux besoins d'apprenants qui sont déjà sur le marché du travail.

L'autre critère à prendre en considération concernait les différences de niveau de langue entre les apprenants. Les uns, le baccalauréat professionnel fraîchement décroché, souhaitaient simplement apprendre quelques expressions appartenant à un champ lexical spécifique. Les autres, pour qui les études an anglais ne représentaient qu'un souvenir lointain, devaient se réapproprier les bases en anglais général en même temps que d'acquérir un vocabulaire spécialisé. Afin de répondre aux attentes de

ce public hétérogène, notre ambition fut d'élaborer un module ayant pour double objectif une remise à niveau et un approfondissement des connaissances.

Nous devions également tenir compte du degré de motivation et d'engagement des participants. Le taux d'abandon étant particulièrement élevé en enseignement à distance, nous souhaitions trouver un moyen de captiver l'attention de l'apprenant tout au long du module. La vidéo nous a semblé la plus appropriée pour remplir ce rôle de motivateur, encore fallait-il l'utiliser à bon escient. Notre effort devait porter à la fois sur le contenu pédagogique et sur la scénarisation.

Après un premier travail consacré à la recherche des outils linguistiques, notamment dans des ouvrages spécialisés en anglais commercial et des affaires, nous souhaitions rédiger des dialogues pouvant susciter l'intérêt des apprenants. Notre public était prioritairement lorrain (LUNO) mais nous envisagions également de l'offrir en ressources complémentaires à des étudiants inscrits en Licence d'anglais au CTU. Nous souhaitions également apporter une touche d'humour au scénario. C'est ainsi que prit forme la société LMD (Lorraine Mirabelles Drinks), fabricant de liqueurs, dont les exportations vers le Royaume-Uni sont menacées. L'ordre du jour de la réunion sera essentiellement consacré à l'analyse de la situation financière et aux mesures à prendre pour redresser l'entreprise.

Le module « Chairing a meeting » devait être utilisé dans le cadre de LUNO avec un tutorat réduit (autoformation complète ou accompagnement limité). Il s'agissait donc d'un apprentissage collaboratif plutôt que d'un apprentissage coopératif. La collaboration (apprenant-machine) demandant plus d'autonomie que la coopération (apprenant-enseignant et machine), nous souhaitions simplifier au maximum l'ergonomie du produit afin de ne pas accentuer les difficultés d'orientation de l'apprenant dans le module. Nous devions nous efforcer d'adapter la technique aux exigences de la pédagogie et non l'inverse.

L'enregistrement vidéo a été effectué par la société de production VIDEOSCOP (Nancy 2) avec des moyens techniques et humains importants. Cinq enseignants anglophones ont joué les différents rôles (directeur, chef des ventes, chef du personnel, représentant du personnel et responsable des ventes au Royaume-Uni).

Nous avons décidé de découper l'enregistrement en 11 séquences afin que l'apprenant puisse adapter l'utilisation des outils à son rythme de travail. Pour chacune des séquences, l'interaction est privilégiée : possibilité de revoir un passage de l'enregistrement, de faire une pause le temps de prendre des notes, de lire la transcription des dialogues, etc.

Mais l'originalité du module se situe au niveau du découpage « chirurgicale » de très courtes séquences. L'apprenant a la possibilité de visionner uniquement les phrases propres à la « conduite de réunion ». Isolées de leur contexte (les problèmes rencontrées par la société LMD...), elles sont données à l'état brut pour inviter l'apprenant à les réutiliser dans un contexte différent. Elles sont accompagnées par des phrases synonymes, disponibles sous forme d'enregistrements audio, afin de montrer l'étendu des outils linguistiques qui servent à prendre ou redonner la parole, reformuler, conclure, etc.

Des outils d'auto-évaluation permettent à l'apprenant de se situer. La variété des tâches, dont certaines ont un caractère ludique, évite la passivité de l'apprenant. Certains exercices comportent plusieurs niveaux de difficulté pour essayer de répondre aux besoins de chaque apprenant. Enfin, un lexique, français-anglais et anglais-français, est également disponible. Nous avons fait le choix d'obliger l'apprenant à cliquer sur un mot s'il veut en obtenir la traduction, afin de le maintenir réceptif.

Une évaluation du module est en cours. Il s'agira notamment de vérifier la corrélation entre le profil des apprenants et leur degré d'autonomie. Mais la priorité de ce questionnaire sera de dresser un bilan sur l'utilisation de la vidéo, ceci dans le but d'améliorer le module existant et d'en tenir compte dans la création de nouveaux produits.

Après la présentation du module « Chairing a meeting », le colloque sera l'occasion d'analyser cette évaluation, en mettant l'accent sur l'apport pédagogique de la vidéo dans la formation.

References

Asensio, M. Strom, J. Young, C. (2001) *Click and Go Video***.** 8th EDINEB Conference 'Educational Innovation in Economics and Business Administration', Nice, June 2001

Banks S., McConnell D., Hodgson V. Author, B. (2004). *Advances in Research on Networked Learning***.** Kluwer Academic Publishers.

Dewald, B.W.A (2000)**.** "Turning part-time students' feedback into video programs"**.** *Education and training***,** Vol 42, Issue 1.

Dembo M., Lynch R. (2004).*The Relationship Between Self-Regulation and Online Learning in a Blended Learning Context*. http://www.irrodl.org/content/v5.2/lynch-dembo.html

Marx, R and Frost, J. (1998) Toward optimal use of video in management education: examining the evidence. *Journal of Management Development***,** Vol 17, Issue 4.

Sherman J. (2003). *Using Authentic Video in the Language Classroom***.** Cambridge University Press.

The teaching contribution of video in the "Chairing a meeting" module

In response to a demand made by LUNO (Lorraine Open University), the teaching team of the CTU (University Distance Teaching Centre) of Nancy 2 offered to create a module entitled "Chairing a meeting in English / Conduire une réunion en anglais".

LUNO is an open university regrouping the main higher education establishments of Lorraine, which are bound by a charter. The project was mainly funded by the Regional Council of Lorraine in September 2000. The members of LUNO have developed numerous course modules in their respective fields of competence.

The online courses are accessible to any employee or job-seeker in Lorraine. The "Chairing a meeting in English" module is thus aimed at learners in vocational training schemes who wish to follow on-line courses. The specificity of the learners was the first criterion to be taken into account in the preliminary draft. Contrary to academic students, who are extrinsic goal orientated (main objective: to obtain a grade), such learners are more intrinsic goal orientated (main objective: personal challenge) and are consequently more likely to set specific learning goals. However, we endeavoured to arouse the interest of those "non academic" students even more and to adapt the teaching material to the requirements of learners who are already in the labour market.

The other criterion was the learners' differences in language level. Some, who have just obtained their professional baccalauréat, just needed to learn expressions belonging to a specific lexical field. Others, who had finished secondary education a long time ago, wanted to revise the basic language skills before acquiring this specialized vocabulary. Our ambition was to elaborate a module whose flexibility enabled the needs of the broadest and most heterogeneous public possible to be met.

We had also to take into account the degree of motivation and commitment of the participants. The rate of dropout being particularly high in distance teaching, we wished to find a means of captivating the learner's attention throughout the module. Streaming video seemed to us the most appropriate tool to fulfil this role providing that we used it advisedly. We focused our efforts both on the teaching contents and on the storyboard.

After the initial work consisting in looking for linguistic tools, mainly in books specialized in commercial and business English, we wished to write dialogues with the objective of arousing the learner's interest. We planned to give a touch of humour to the content. Thus took form the LMD company (Lorraine Mirabelle Drinks), a liquor manufacturer, whose exports to the United Kingdom are threatened. The agenda of the meeting would primarily be devoted to the analysis of the financial situation and to the measures taken to put the company on its feet again.

The "Chairing a meeting" module would be used within the framework of LUNO with a reduced tutorial scheme (complete self-training or reduced tutorial). It was a collaborative training rather than a co-operative one. Collaboration (learner-machine) asking for more autonomy than co-operation (learner/teacher and machine), we wished to simplify ergonomy (video and audio quality, screen visibility, design, weight) to offer the best learning environment possible and to facilitate the learner's orientation within the module. We had to adapt the technique to the requirements of pedagogy and not the reverse.

The video recording was conducted by VIDEOSCOP, a production company of Nancy 2, with significant technical means and staffing. Five native speakers played the various parts: managing director, personnel manager, sales manager, managing director, staff representative, salesperson for Britain.

We decided to divide the recording into 11 sequences to allow the learner to adapt his use of the tools to his own pace. For each sequence, we privileged interaction: the possibility of reviewing a passage of the recording, of making a pause, etc.

The originality of the module lies in the "surgical cutting" of very short sequences. It offers the learner the possibility to view only the sentences useful to "chair a meeting" .Isolated from their context (the problems encountered by the LMD company...), learners are invited to use them again in a different context. They are accompanied by synonymous sentences, available in the form of audio recordings, in order to offer a wide range of linguistic tools that are used for taking or giving the floor, reformulating, concluding, etc.

Different learning activities provide diverse forms of formative assessment. The variety of short activities encourages the learner. Some exercises comprise several levels of difficulty to try to meet the needs of each learner.

Lastly, a lexicon (French-English and English-French) is also available. We made the choice to oblige the learner to click on a word if he wants to obtain the translation of it, in order to enhance his receptiveness.

An evaluation of the module by the learners has been carried out. The priority of this questionnaire is to evaluate the use of video. The form includes a rating scale and open-ended questions that ask what the learner thinks about the training materials and activities. This research does not include the evaluation of the trainer and the training environment. Our main objective is the upgrading of new video based material that we want to realise. We also intend to revise the different tasks offered in this module to fit the needs of future learners.

After the presentation of the "Chairing a meeting" module, the conference will be the occasion to analyze the results of this evaluation, by stressing the contribution of video in the module.

The Digital Dinneen Project : Further Avenues for CELT.

Julianne NYHAN

CELT Project, History Department, University College Cork, Ireland.

Rev. Patrick Stephen Dinneen's Irish-English Dictionary is widely regarded as the most authoritative scholarly dictionary of modern Literary Irish currently available. Published in 1934, it has seen numerous reprints, but is not available in digital form. Indeed, no scholarly Irish-English dictionary of modern Literary Irish is available in digital form.

The Corpus of Electronic Texts, supported by the Irish Research Council for the Humanities and Social Sciences, has embarked upon a three year project to deliver a digital version of Dinneen's Dictionary. The Research Associate responsible for the preparation of the Digital Dinneen is Julianne Nyhan, and the Principal Investigator on the project is Professor Donnchadh Ó Corráin, University College Cork (hereafter UCC). Dr Gregory Toner, from the University of Ulster at Coleraine, and Dr Seán Ua Súilleabháin from the Department of Modern Irish at UCC, are associate investigators.

Dinneen's dictionary is a complicated document that contains a wealth of information. Its data comprises, inter alia, headwords, grammatical information, definitions, usage examples and translations, as well as references to dialectical sources and to informants used by Dinneen. This paper will briefly discuss the TEI mark-up of other Irish language dictionaries, such as the eDIL (electronic Dictionary of the Irish language). The use of TEI to encode Dinneen will then be illustrated, and the necessary customisations briefly discussed.

In addition to creating a digital edition of Dinneen, the research assistant is endeavouring to develop an edition that is more user-friendly that the hard copy edition of the same work. Much of the information contained in the dictionary remains inaccessible even to experienced speakers of the language, because the hard copy contains mixed font (Cló Gaelach/Roman), and many people

are not able to read Cló Gaelach. This barrier will be removed from the digital edition, and end-users will have the choice of viewing the text in either mixed font or Roman font only. Furthermore, the dictionary contains orthographic forms that pre-date the spelling reform of 1946, and as many current day speakers are unfamiliar with such orthography, they have difficult locating headwords. The systematic incorporation of modern orthographic forms as meta-data will enable access for modern speakers who are unfamiliar with historical spelling - and today they are probably the majority of Irish speakers.

This paper will focus on how the Digital Dinneen will be integrated into the existing CELT infrastructure, and into the wider infrastructure of Humanities Computing in Irish. An electronic Lexicon of variants of medieval Irish, the subject of Julianne Nyhan's PhD research, will become available on the CELT website in 2006/2007. Incorporating id references into Dinneen will allow end users to trace a word back to its medieval form in the Lexicon of variants. Furthermore, for the last two years, CELT has been involved in a cross-border collaboration with the University of Ulster at Coleraine, where an electronic edition of the *Dictionary of the Irish Language* (eDIL) is being prepared. It is hoped that links between Dinneen and the eDIL will be easily generated. The research carried out at CELT into an XSLT lookup tool to facilitate links between works cited in eDIL, and the corresponding text in the CELT corpus, will also be extended to Dinneen. This mechanism will enable scholars on-line access to works cited by Dinneen - many of which are available only in the best research libraries. A Javascript plug-in, that enables end users of the CELT corpus to highlight a word, click on a lookup icon and retrieve the word in question from either the Lexicon of medieval Irish or the Digital Dinneen will also be discussed. Finally, the possibilities for future research will be touched upon, for example, the possibility of augmenting Dinneen's citations with examples from the CELT corpus as more texts are added to it.

Monumenta Frisingensia — a Digital Edition of the Oldest Slovenian Text

Tomaž ERJAVEC
Institute of Slovenian Literature, Scientific Research Centre

Matija OGRIN
Department of Knowledge Technologies, Jožef Stefan Institute

Monumenta frisingensia – the 10th century Freising Manuscripts (Slovenian: Brižinski spomeniki) – are the oldest written Slovenian text and also the oldest Slavic text written in the Latin alphabet.

The manuscripts consist of three religious texts, comprised within the Codex latinus monacensis 6426 (Munich State Library). The Monumenta frisingensia (MF) were written in a script based on Carolingian minuscule after AD 972 (MF II and MF III probably before 1000) and before 1039 (MF I most probably by 1022 or 1023). The provenance of the MF is either Upper Carinthia or Freising (now in Austria or Germany respectively); in any case, they were used in the Carinthian estates of the Freising diocese.

Slovenian historians, linguistic and literary scholars consider the MF to be the very first document of early Slovenian language. This is the reason why these manu scripts bear an outstanding importance not only for the scientific comprehension of the development of Slovenian language and literature but also for Slovenian national identity and historical consciousness. In this respect Monumenta frisingensia represent the genesis and the beginnings of Slovenian national individuality.

The digital edition of the Monumenta is based upon a printed one (the critical edition by the Slovenian Academy of Sciences and Arts), which encompasses a complex apparatus with facsimile, diplomatic, critical and phonetic transcriptions, translations into Latin and five modern European languages, and a dictionary covering these transcriptions and translations (plus Old Church Slavonic). This edition furthermore contains

introductory studies, apparatus, a bibliography and an index of names.

The MF are preserved in one single manuscript only. For this reason, the traditional philological and critical task – to examine the ms. witnesses – is not possible in this case. On the contrary, it is of great importance that the edition is founded on the opposite concept: on that of reception. In what way the MF were studied, understood and published? What were the main differences between the transcriptions and editions? Can a digital edition grasp at least crucial points of their research during the two centuries from their discovery, 1806–2006?

For this reason, besides all the components of the printed edition, the digital one contains a selection of earlier diplomatic and critical transcriptions (in full text) and enables parallel views to compare them. This is of considerable importance as the Frisingensia were often a subject of polemical scientific interpretations, especially with regard to the place of their origin and their linguistic genesis. The edition should, therefore, enable a pluralism of readings and understandings as well as a hierarchy of proofs and ascertainments. By inclusion of all major transcriptions of the Freising Manuscripts since the very first diplomatic transcription in 1827 (P. Koeppen), the e-edition offers a panoramic retrospective of 200 years of research of these precious manuscripts. Such a retrospective, evidenced by historical transcriptions themselves, can reveal cultural and philological history of the Frisingensia.

As another extension, specific for digital edition, we decided to include the digitized audio recordings of the read manuscripts and integrate the segmented audio files to both phonetic transcription and modern Slovenian normalization. The scope of this audio extension is to offer a reconstruction (based on historical linguistics) of medieval Slovenian speech which can be sensibly juxtaposed and compared with transcriptions. We plan to publish a pedagogical adaptation of this e-edition as well, aimed for school-reading, where audio component with the archaic pronunciation can produce a clear image of historical development of the Slovenian language.

Such materials present significant challenges for encoding, esp. the high density and variety of markup, extreme parallelism (per-line alignment between the text views), and special historic and phonetic characters used in the transcriptions.

The paper details our methodology used to turn the printed edition (plus extensions) into a Web edition with a standardized encoding, extensive hyperlinking, and multimedia capabilities. We concentrate on the following issues:

- Structuring of materials in a text editor
- Adoption of standard solutions: Unicode/XML/ XSLT/TEI-P4/
- Collaborative practices using fast prototyping and cyclical improvement: up-converting to XML–TEI --> down-converting to HTML --> proofs, corrections --> re-applying the up-conversion …

Our presentation details these issues, and highlights the most challenging aspects of entire process, where the central goal is to enable a complex and lively communication with the 10th century Monumenta frisingensia.

Acknowledgments

The Monumenta Frisingensia e-edition was prepared (and is freely available on-line) within a collaborative project “Scholarly digital editions of Slovenian literature” (http://nl.ijs.si/e-zrc/index-en.html). It is a joint project of the Institute of Slovenian literature (Scientific Research Centre of the Slovenian Academy of Sciences and Arts, Ljubljana) and the Department of Knowledge Technologies (Jožef Stefan Institute, Ljubljana). The project was undertaken to provide an application of traditional text-critical principles and editorial technique to the publication of selected Slovenian texts in the digital medium. Among these editions, the current digital edition of the MF occupies a superior position.

References

Bernik F. et al., ed. by (2004). Brižinski spomeniki. Monumenta frisingensia. Znanstvenokritièna izdaja. Založba ZRC, Ljubljana.

"The Margins of the Scholarship: Children's Literature and the Hypertext Edition"

Elan PAULSON

English - University of Western Ontario

Based on a consideration of recent editorial criticism concerning issues of marginality, this presentation calls for developing a greater number of hypertext editions of children's literature. To demonstrate an engagement with and a response to such criticism, this presentation introduces a hypertext edition of the literary tale "The Grey Wolf" by George MacDonald (1871), in which the guiding methodology and digital design, interface, materials, and critical apparatus fit the particular research and educational needs of scholars and students of children's literature.

This presentation proposes that more critical hypertext editions of children's literature will continue to help reduce scholarly assumptions that dismiss children's literature as straightforward, transparent, and marginal to the more central concerns of literary criticism. Recent feminist and minority textual studies critics have already identified some of the ways in which editorial and publishing practices reproduce hegemonic and positivist discourses that relegate certain authors and genres to a marginal status in bibliographical and literary studies. For example, feminist critic Ann Thompson argues that the masculinist discourses implicit in editorial conventions have produced critical editions that re-inscribe male-biased assumptions, selections, omissions, and interpretations of texts. Similarly, William Andrews points out that while the literature of minority authors is beginning to appear more frequently in electronic archives and student editions, this editorial work is done in place of developing research-based critical editions for the use of professional academics. But although many kinds of children's literature infrequently contain textual allusions, challenging language, and references that require annotative explanations, children's literature scholar Jack Zipes argues that the simplistic narratives and symbolic elements that characterize children's literature often carry the most potent and subversive political agendas for the very reason that they are considered harmless and transparent. Taking cues from these critics' awareness of how the elision of marginal voices and texts creates disproportionate literary interpretation, dissemination, and canonization, this presentation considers the dynamic potential of communicating the relatively marginalized genre of children's literature in critical hypertext editions in ways that would benefit scholars and students. With each new publication of a text that belongs to a "marginal" genre, hypertext critical editions produce what D.C. Greetham describes in *Theories of the Text* (1993) as the "deferral and dispersion" (5) of the pervasive hierarchical and hegemonic editorial conventions that continue to privilege certain texts, authors, and genres over others.

Thus, this presentation aims to demonstrate the theoretical, practical, and pedagogical fit between research strategies in children's literature studies and the capabilities of hypertext to present and connect visual, contextual, literary and bibliographical material. Because the oral folk tale has no single origin, and because critical work in the area of children's literature often compares thematics and narrative structures in related historical, national, and transnational texts, hypertext editions of children's literature can be designed to present multiple versions of narrative material in ways that de-centre traditional editorial principles of "best text" and "authorial intention." Moreover, critics have recently argued that the editor's annotations contribute to the (re)production of the text in fundamental ways, further problematizing the notion of "authorial intention." For example, in discussing some of the methodological issues that the editors faced in creating *The Prufrock Papers: A Hypertext Resource for "The Love Song of J. Alfred Prufrock"* (1999), Peter Stoicheff and his colleagues regard the hypertext edition as that which "permits a variety of copy-texts" whose hyperlinks and critical apparatus create "a new kind of interpretive 'copy-text'." Feminist editor Brenda Silver further proposes that the adaptation of a text is a form of editorializing that does more than interpret an original text; adaptation, in the author's view, performs its interpretation. In a related vein, Julia Flanders argues that the xml transformation of the codex text into hypertext necessitates semantic tagging, and that editor's interpretive annotations combine with the text at the xml level. It follows, then, that any hypertext edition

is not only a re-presentation of the author's text but is itself an adaptation, in which the author's content and the editor's presentation of the content join as a single, collaborative performance. Guided by this editorial criticism, then, an annotated hypertext edition of a children's literary tale would enact the editorial theory that encourages a de-centering and de-privileging of traditional editorial conventions by offering a maximum amount of relational play between textual versions and literary interpretations, of which the hypertext edition *itself* is one.

Not only does the editorial theory concerning issues of marginality encourage a de-centring of versions and authorial intent, but hypertext editions that prescribe to such methodologies can also offer the user diverse interpretive materials for analyzing children's literature. For example, literary tale scholars often trace the historical development and codification of oral folktales, collate and compare versions, and examine codex editions as cultural artifacts in that they reflect and reproduce political, economic, and moral norms and values. Also, the hypertext edition can offer additional interpretive dimensions to the text by reproducing bibliographical information in digital images, including watermarks and signatures, as well as graphics such as the scripts, borders and illustrations. The hypertext edition can also present audio or visual cinematic adaptations of the literary tale for comparison. Thus, there is immense potential for reproducing children's literature in hypertext format, particularly since the electronic text may – with hyperlinks, split screens, digital image rendering and audio/visual clips – present a variety of research materials relevant to children's literature studies that the traditional codex edition could not. In the spirit of Peter Shillingsburg's assertion that "experimentation and a variety of approaches are to be encouraged" in hypertext edition production, this presentation suggests ways in which the digital transmission of children's literature de-centres editorial conventions while facilitating scholarly research in this area.

To demonstrate this fit between certain editorial theories concerning marginality and adaptation and the practice of studying and comparing textual and literary elements of children's literature, this presentation introduces a hypertext edition of "The Grey Wolf," first published in 1871 in a ten-volume set of MacDonald's work, entitled *Works of Fancy and Imagination*. This hypertext edition offers different interpretations of the text through colour-coded popup annotations and through the presentation of related texts in a split-frame interface. It also contains an enumerative and analytical bibliography of the 1871 edition, and compares this edition to a paperback edition published by Eerdmans over one hundred years later (1980). Furthermore, this bibliographical component of the edition familiarizes student users with the field of textual studies, and suggests how bibliographical elements may contribute to literary interpretations in meaningful ways. My approach to editing "The Grey Wolf: A Hypertext Edition" is guided by an integrated matrix of textual studies theory and literary criticism, an approach that has been called for by a growing number of textual studies and literary critics, such as Michael Groden and Marie-Laure Ryan. This approach also draws attention to the issues of marginality, versioning and the hypertext apparatus through which the text is mediated, as well as how this remediation inevitably influences the reception and interpretation of the text.

References

Andrews, W. (1997). "Editing Minority Texts." *The Margins of the Text*. Ed. by D.C. Greetham. Ann Arbor: U of Michigan P. 45-56.

Flanders, J. (1997). "The Body Encoded: Questions of Gender and the Electronic Text." *Electronic Text: Investigations in Method and Theory.* Ed. Kathryn Sutherland. Oxford: Clarendon, 127-43.

Greetham, D.C. (1999) "Gender in the Text." *Theories of the Text*. Oxford: Oxford UP, 1999. 433-486.

Groden, M. (1991). "Contemporary Textual and Literary Theory." *Representing Modernist Texts: Editing as Interpretation*. Ann Arbour: U of Michigan P.

MacDonald, G. (1871). *Works of Fancy and Imagination.* Vol.10. London: Strahan and Co., 1871.

---. (1980). *The Gray Wolf and Other Stories*. Grand Rapids, MI: Eerdmans.

Ryan, M-L. (1999). *Cyberspace Texuality: Computer Technology and Literary Theory*. Bloomington:

Indiana UP.

Shillingsburg, P. (1997). "Guidelines for Electronic Scholarly Editions." *Modern Languages Association of America Committee on Scholarly Editions*. 12 Oct 04. http://sunsite.berkley.edu/MLA/guidelines.html

Silver, B. (1997). "Whose Room of Orlando's Own? The Politics of Adaptation." *The Margins of the Text*. Ed. D.C. Greetham. Ann Arbor: U of Michigan P. 57-82.

Stoicheff, P. et al. (1999). "The Editor in the Machine: Theoretical and Editorial Issues Raised by the Design of an HTML Literary Hypertext." *The Prufrock Papers: A Hypertext Resource for "The Love Song of J. Alfred Prufrock."* University of Saskatchewan. 06 Jan 2005. http://www.usask.ca/english/prufrock/index.html

Thompson, A. (1997). "Feminist Theory and the Editing of Shakespeare: The Taming of the Shrew Revisited." *The Margins of the Text*. Ann Arbor: U of Michigan P. 83-104.

Zipes, J. (1983). *Fairy Tales and the Art of Subversion*. New York, NY: Wildman, 1983.

Just Different Layers? Stylesheets and Digital Edition Methodology

Elena PIERAZZO

University of Pisa - Department of Italian Studies

In the last few years a consistent number of papers focussed on new theoretical frameworks for scholarly digital editions practice (Vanhoutte 2000a, Vanhoutte 2000b and Flanders 1998). The main attention has been paid on text encoding, on the production of *apparatus criticus or variourum* and its transposition form the paper to the digital format (Vanhoutte 2000a, Vanhoutte 2000b, McGann 1996 and Lavagnino 1996). Another focus is the role of the editor and whether a scholar edition must have as a goal the production of a text (a quotable text) or just different textual materials in order to allow the reader to choose his/her version of the text; a third possibility being an equilibrium (to be found) between the two school of thinking (Vanhoutte 2005). Connected to the latter point is the consideration that a digital edition based on encoding can be re-used for many purposes (e.g. electronic inspection, computer assisted analysis), and gives the possibility of producing different versions of the same text (critical, diplomatic, facsimile, reading editions, hypertext editions), according with different kind of users and readers from the same encoded text by the application of different stylesheets. In this paper I will address to a different consideration of stylesheets to be seen not only as a tool to produce different layers of the encoded texts, but as an essential component of the scholar's work and a possible shortcut toward a compromise in the dichotomy textual criticism vs. cultural criticism, as in Vanhoutte 2005.

Some time ago I started to prepare a digital edition of an Italian Renaissance's comedy, *Lo Stufaiuolo*, written by Anton Francesco Doni, presumably in 1550-1551. We posses two different autographic manuscripts of the text, each of them witnessing a different and remarkable version, a fact that involves the necessity of a synoptic edition. The first question to be answered in digital

edition field is certainly which kind of edition to produce: an editor-oriented or a user-oriented edition. In my opinion, according with the Italian editorial school's theories, an editor **must** produce a quotable, reliable text or at least, to provide an orientation to the reader: the editor is probably the person who know the most about the text, and can not skip the task to provide the readers his circumspect opinion/version of the text. On the other side, an editor **must** also provide all the documentation possible of his work, according to a fundamental presumption of the scientific nature of a critical edition that lays on the possibility for the reader to reproduce/control/verify the editorial work. For that reasons I chose to provide different versions of the text (of both texts, in this case), ranging from a minimum to a maximum of editorial intervention, in order to satisfy different readers' needs, accompanied by ancillary documentation on the editorial work and by the photographic reproduction of the source.

Due to the big interest of the Doni's language, a peculiar point of the edition would have been the preserving of the original face of the text, allowing at the same time to non-scholar readers to appreciate the text itself, that is very funny and with a huge literary value. Starting the work, I focussed my attention to the very first level of the language: orthography and punctuation. In paper-based critical editions there is often the necessity of re-write the text in order to allow modern reader to easy decrypt it. The orthography and punctuation rules of most modern languages have sensibly changed from Middle Age since today. At the dawn of the writing of modern languages we can not even speak of orthography but of different writing habits, relayed to different cultural centres. Furthermore, very early texts are in *scriptio continua* and for many centuries the word-borders have been sensibly different from the present ones. Orthography and punctuation have reached the present relative stability through different phases and approximation levels in relative recent years; for Romance languages, for example, the Latin graphical habits had a long persistence in writing. Normally, in the editorial practice, editors "translate" from ancient writing-face, according to modern rules, declaring in preliminary sections of the printed book the transcription's criteria. In this way paper-based editions are able to give just a summary of the original's reality and can not be used by scholars who want to study the original language face, punctuation, orthography or the writing supports: as for *apparatus criticus or variorum*, transcription criteria in prefatory pages do not allows *de facto* the possibility of reproducing the editorial work. In such cases to turn back to original or facsimile editions is the only possibility, but such editions are totally resistant to electronic inspections and limit the scholar's chance of availing large corpora of data.

This limitation can be overruled by electronic editions based on text encoding. In fact, the encoding allows adding new meaningful layers without losing any aspect of the original face of the text. For example, adopting the TEI encoding schema, it is possible to encode the original face of the text as follow:

> <orig reg="Perché">Per che</orig> le lettere sono state sempre poste <orig reg="nei">ne i</orig> degni <orig reg="e">et</orig> <orig reg="onorati">honorati </orig> luoghi…

We can also preview a more sophisticated encoding able to classify the kind of regularization provided (after having extended the TEI encoding schema in order to enable a "type" attribute for the <orig> element):

> <orig reg="Perché" type="wordBorder">Per che</orig> le lettere sono state sempre poste <orig reg='nei' type="wordBorder">ne i</orig> degni <orig reg='e' type="lat">et</orig> <orig reg='onorati' type="lat">honorati</orig> luoghi…

With such encoding it will be possible to display the text according to the <orig> face or to the reg face using suitable XSLT stylesheets. In case of multiple editions from the same text, we can also think to editions regularizing each time particular sets of "type", according with different purposes, for example considering just punctuation or just orthography regularization.

The usage of stylesheets can be also anticipated from the delivery phase to the encoding/editorial phase. A useful application of stylesheets during the encoding can be managed by the encoder/editor to monitor its own work and its interventions on the text. Stylesheets, in fact, are able to produce at any moment of the work statistical data in real time, a fact which can help the editor to better understand the writing habits of the author or of the copyist and, at the same time, to keep under control its own work. Similar considerations can be made over the encoding of readings and lemmas, where lists of refused or accepted readings can profitably support the

editorial work in every phase. Stylesheets are a low-cost (or a non-cost), flexible and suitable technology, relatively easy to understand in its principal applications and can be managed directly by the editor, not a secondary concern in low budget projects.

In the delivery phase, the same lists and data can be produced to the scientific community, giving explicitness to the editorial practice and overruling the compactness of the introductory notes of paper-based editions.

Stylesheets can be reasonably considered not only an ancillary part of the editorial practice, a technology to be recalled **after** the finish of the scholar work (maybe committed to computer science people), but a part of the work committed to the editor himself/herself, able to support his/her tasks in monitoring the work step by step and able to produce different texts for different tastes, under the editorial control.

The paper will conclude by presenting the digital edition of the Doni's *Stufaiuolo*, in order to exemplify the results obtained with such methodological approach.

References

Flanders, J. (1998). "Trusting the Electronic Edition," *Computers and the Humanities* 31/4.

Lavagnino, J. (1996). "Reading, Scholarship, and Hypertext Editions", *TEXT*, 8

McGann, J. (1996). "The Rationale of HyperText", *TEXT*, 9.

Sperberg-Mcqueen, C.M. and Burnard, L. (2002). *TEI P4 Guidelines for Electronic Text Encoding and Interchange: XML-compatible Edition*. (P4). Oxford, Providence, Charlottesville, & Bergen: The TEI Consortium. Available also at http://www.tei-c.org/P4X/.

Vanhoutte, E. (2000a). "Textual Variation, Electronic Editions and Hypertext". Paper. ACH/ALLC 2000. Glasgow (UK): Glasgow University. 23 July 2000. Abstract available at: http://www2.arts.gla.ac.uk/allcach2k/Programme/session4.html#433.

Vanhoutte, E. (2000b). "Where is the editor? Resistance in the Creation of an Electronic Critical Edition", *DRH98: Selected Papers from Digital Resources for the Humanities 1998,* ed. Marilyn Deegan, Jean Anderson, and Harold Short (London: Office for Humanities Communication, 2000).

Vanhoutte, E. (2005). "Editorial theory in crisis. The concepts of base text, edited text, textual apparatus and variant in electronic editions". Paper. DRH 2005. Lancaster (UK): Lancaster University. 5 September 2005.

Vanhoutte E. and Van der Branden, R. (2005). "Describing, Transcribing, Encoding, and Editing Modern Correspondence Material: a Textbase Approach". Fred Unwalla & Peter Shillingsburg (eds.), Computing the edition. Toronto, Toronto University Press. Preprint on http://www.kantl.be/ctb/pub/2004/comedvanvanfig.pdf.

De l'index nominum à l'ontologie. Comment mettre en lumière les réseaux sociaux dans les corpus historiques numériques ?

Gautier POUPEAU

École Nationale Des Chartes

Dresser les index fait partie intégrante du travail d'élaboration d'une édition critique de sources. Dans le cadre du support papier, les index permettent de donner aux lecteurs une idée du contenu et, surtout, d'accéder précisément à une information : un nom de lieu ou de personne dans le cas des index nominum, ou un concept dans le cas des index rerum. L'index est en revanche quasiment absent des éditions numériques et des bases de données textuelles, et plusieurs raisons peuvent expliquer cette absence. D'une part, dans la plupart des cas, le but de telles entreprises n'est pas de proposer des éditions critiques de sources, mais plutôt des bases de données textuelles. Les formulaires d'interrogation constituent le point d'accès principal à l'information, et la primauté du texte intégral a pu faire croire dans un premier temps à l'inutilité de l'index. D'autre part, mettre en place un index représente un travail long et fastidieux, qui entraîne un coût non négligeable difficile à supporter pour des éditeurs commerciaux.

Pourtant, même dans le cadre du support numérique, en particulier sur le Web, la pratique de l'indexation trouve des justifications. Ainsi, alors que dans le cas de corpus médiévaux, la lemmatisation automatique est très difficile à mettre en place, surtout sur les noms de personnes et de lieux qui présentent les graphies les plus disparates, l'indexation peut constituer une première réponse à ce problème. Par ailleurs, l'index offre un panorama des noms de lieux et de personnes présents dans l'ouvrage, ce qui est impossible avec les bases de données textuelles qui ne proposent aucun moyen de prendre connaissance du contenu du corpus de manière globale. Aujourd'hui commencent à apparaître sur le Web des éditions critiques proprement dites, qui s'appuient sur la structure originelle de ce type de travail de recherche ; logiquement, elles comprennent aussi un index, celui-ci faisant partie intégrante de la tradition scientifique de l'édition critique. C'est en adoptant cette démarche que se dégagent les prémices d'une exploitation innovante de l'index dans le contexte de l'édition électronique.

Dans le cadre de l'organisation hypertextuelle de l'information, alors que l'index n'était qu'un outil de repérage souvent difficile d'emploi sur le support papier, il devient un moyen rapide d'accès à l'information surtout dans le cas d'un survol et d'une prise de contact du corpus par l'internaute, et fait apparaître de nouveaux parcours de lecture.

Malgré ces disparités existantes d'un ouvrage à l'autre, quatre parties récurrentes peuvent être distinguées dans un index :

- L'entrée de l'index ;
- L'ensemble des formes présentes dans le texte et faisant référence à cette entrée ;
- Des indications biographiques et/ou généalogiques ;
- L'emplacement dans l'ouvrage des occurrences de l'index.

Les index de noms sont souvent organisés de manière hiérarchique, induisant des liens entre les entrées. Toutes ces informations constituent une première couche interprétative. Pourtant, l'index est cantonné dans ce rôle de repérage et d'accès à l'information, et ces données sont rarement exploitées pour traiter le corpus. Or, ce travail pourrait servir de socle à la construction d'une ontologie, qui constituerait un index commun à plusieurs corpus, jouant en quelque sorte le rôle d'un fichier d'autorité.

Une ontologie informatique, à ne pas confondre avec celle des philosophes, permet de définir des concepts et de décrire les relations qui peuvent exister entre ces différents concepts. L'avantage d'une ontologie sur une base de données relationnelle réside dans la possibilité de définir des règles logiques entre les concepts et entre les relations, et de dépasser ainsi le concept d'organisation hiérarchique qui prévaut dans les formes traditionnelles de l'index. Par exemple, soit la relation « frère de » symétrique et la propriété « A frère de B », alors l'inférence « B frère de A » est déduite automatiquement. De la même façon, soit la relation « enfant de » transitive de la relation « parent de » et la propriété « A parent de B », alors la propriété « B est enfant de A » est déduite automatiquement. Ce mécanisme, qui peut sembler simple et logique à

l'appréhension humaine, est en réalité assez complexe à reproduire dans un contexte d'automatisation du traitement de l'information.

Il existe plusieurs langages pour mettre au point une ontologie. Le langage OWL défini par le W3C et basé sur RDF, permet d'envisager, grâce à sa syntaxe XML, une transformation simple par feuilles de style XSL des données du corpus encodés en XML vers l'ontologie au format OWL. Pour peupler l'ontologie, en récupérant à la fois l'ensemble des noms de personnes et de lieux indexés sous une forme régularisée, et la relation de ces noms avec les différents documents du corpus, nous pouvons alors nous appuyer sur la structuration en XML des corpus historiques selon la DTD TEI.

Pour un corpus encodé selon la version dite P4 de la TEI, l'élément vide <index/> peut être utilisé pour indexer un point dans le texte. Mais cet élément ne permet pas la récupération des différentes formes des entrées indexées. En revanche, l'élément <persName> accompagné de l'attribut reg permet d'encoder un nom de personne et d'indiquer une forme régularisée, et l'élément <placeName> permet quant à lui d'encoder un nom de lieu.

Une fois le cadre structurel des entrées d'index ainsi défini, il importe de mettre en place ce qui sera la structure de l'ontologie adaptée pour représenter les réseaux sociaux. L'indication de l'emplacement des occurrences d'une entrée d'index dans le corpus structuré en XML permet de créer automatiquement les relations entre les noms de personnes ou de lieux et les unités structurelles du corpus. Pour autant, le rôle ou la place des personnes dans l'unité structurelle peuvent être précisés. Dans le cas de chartes, nous pouvons préciser s'il s'agit de l'auteur de l'acte juridique, du bénéficiaire ou d'un témoin, par exemple. Pour finir, il faut ajouter les relations entre les personnes et entre les personnes et les lieux, en s'appuyant sur les relations mises en lumière à travers la présence dans un document et/ou sur d'autres sources de première ou de seconde main.

Trois méthodes de visualisation sont à mettre en œuvre pour exploiter pleinement le potentiel de cette démarche. Tout d'abord, une visualisation proche de la mise en page d'un index traditionnel est indispensable pour assurer le rôle de transition vers l'appropriation du support électronique par les chercheurs. Dans un second temps, l'ontologie est proposée dans une forme apte à donner la vision d'ensemble du contenu : des interfaces de navigation à facettes offrent sous forme de listes combinables des parcours dans le corpus qui sont entièrement définis par l'utilisateur. Enfin, une visualisation graphique révèle pleinement le potentiel de l'ontologie. Le graphe fait apparaître de manière tangible les relations entre les personnes, les documents et les lieux, ainsi que le type de relations qui ont été définies. C'est alors que peuvent être mis en lumière les réseaux sociaux qui existent à l'état sous-jacent dans les documents.

Grâce à cette dernière méthode de visualisation sous forme de graphe, l'ontologie prouve son utilité dans le cadre de l'encodage des index nominum. Une base de données relationnelle, étant incapable de modéliser des relations autres que hiérarchiques, ou de déduire les inférences des relations exprimées, rend impossible la modélisation complète des réseaux sociaux. L'ontologie, au contraire, autorise ce processus et permet à l'index de dépasser son rôle de point d'accès pour devenir un véritable outil d'analyse de corpus. Les réseaux sociaux mis en lumière par l'ontologie sous forme de graphe deviennent plus faciles à appréhender, ce qui constituera, une fois cette technologie déployée sur des corpus historiques significatifs, un moyen essentiel d'étudier la présence d'un groupe de personnes dans un espace géographique en fonction de leurs relations. La mise en place de l'index nominum sous forme d'ontologie, s'appuyant sur des corpus structurés en XML, montre bien comment une technologie comme RDF, en transformant la façon dont l'information est modélisée, peut décupler l'intérêt d'un outil bien identifié.

The elaboration of critical edition of sources basically includes the setting up of indexes. In the paper media, the index provides an overview of the content, and makes it possible to access directly a piece of information : the name of a place or a person in the case of index nominum, or a concept in the case of index rerum. Though, indexes are hardly ever integrated in electronic publishing and textual databases, and this can be easily explained. On the first hand, in most cases, the goal of such initiatives isn't to propose critical editions of sources, but to gather vast amounts of texts. Search forms are the main access tool, and the primacy

of full text used to make indexes seem useless. On the other hand, setting up an index is a long and fastidious work, and the cost is not affordable for commercial publishers. Though, even within the digital media, and particularly on the Web, indexing is a worthy effort. It can solve the problem of lemmatisation, which is very hard to proceed automatically on a medieval corpus, especially regarding the diversity of writings of the names of people and places. Moreover, the index provides an overview of the names that appear in the work, when in textual databases there is no way to have a global approach of the corpus. Today, we're beginning to see on the Web real critical editions, that rely upon the original structure of this kind of scholarly work ; logically, they also include an index, since this is part of the academic tradition of critical edition. Through this approach, it is possible to outline innovative uses of indexes within the context of electronic publishing.

While in the paper media the index was only a hard to use location tool, within hypertextual organisation of information, it becomes a quick access tool, gives the overview of the corpus when discovering it for the first time, and reveals new courses of reading. Beside some disparities between different works, an index is composed of four stable parts :

- the entry
- the different writings in the text corresponding to this entry
- biographical and/or genealogical explanations
- the location of occurrences in the text.

Name indexes are often organized hierarchically, inducing links between entries. All this information represents a first interpretative layer. Though, the index is stuck in this purpose of finding and accessing information, and these data aren't used to analyse the corpus. It would be possible to use this work as the basis of an ontology, which would be a common index for various works or corpus, playing a role similar to a list of authorities. A computer ontology, not to be confused with philosophical ontologies, allows to describe concepts and relations between these concepts. The advantage of an ontology over a relational database is to enable logical rules between concepts and relations, and to go further than the hierarchical classification traditionally used for indexes. For instance, the symmetric relation " brother of " and the property " A is brother of B " allows to automatically deduce the inference " B is brother of A ". Similarly, the relation " child of ", transitive to the relation " parent of ", and the property " A is parent of B " the inference " B is child of A " is deduced. This mechanism, for simple and obvious that it seems to a human being, is actually quite complex to reproduce in the context of information treatment automation.

There are various languages to set up an ontology. The OWL language, defined by the W3C and based on RDF, allows, thanks to the XML syntax, a simple XSLT transformation of the data encoded in the corpus in XML towards the OWL ontology. To populate the ontology, we can use the XML structure of historical works encoded following the TEI DTD, to gather all the names of people and places indexed in a regular form, and the relation of these names with the documents in the corpus. For a corpus encoded with the P4 version of TEI, the empty tag <index/> can be used to locate a point in the text, but it is not possible to reuse the different forms of the indexed entries. On the contrary, the <persName> tag with the reg attribute allows to encode the name of a person and to indicate the regular form, and so is it with the <placeName> for places. Once we defined the structural framework on index entries, we can work on the intended structure of the ontology in order to represent social networks. The location of the occurrences of the entry in the corpus structured in XML allows to automatically create relations between names of persons and places, and the structural entities of the corpus. Thus we can also indicate the role or place of the persons : in the case of a charter, we can indicate if it's the author of the juridical action, the beneficiary or a witness, for example. Finally, we have to add the relations between people and placing, basing ourselves upon the relations revealed by the document, or other primary or secondary resources.

Three visualisation methods can be used to fully reveal the benefits of this approach. In the first place, a visualisation close to the presentation of traditional indexes is required in order to ensure the transition towards assimilation of the digital media by the scholars. Second, the ontology is proposed in a manner proper for giving an overview of the content : faceted navigation interfaces propose combined lists that allow the end-user to define by himself different courses in the corpus. Finally, the graphical visualisation reveals the full potential of the ontology. The graph makes the relations between people, documents and places

tangible, as well as the relations types that have been defined. Thus we can unveil the social networks that exist in a hidden state in the documents. The visualisation in the form of a graph proves the usefulness of ontologies when encoding an index nominum. Because a relational database cannot show relations other than hierarchical, or deduce inferences of expressed relations, it cannot completely realise the modelling of social networks. On the contrary, an ontology allows this process and transforms the index from a classical access point into an real tool for analysing the corpus. The social networks revealed by the ontology in the graph form become easier to apprehend, and this will constitute, once this technology is implemented on significant historical corpus, an essential manner to study the presence of a group of people in a geographic space according to their relations. The setting up of the index nominum as an ontology relying on XML structured works shows how a technology like RDF can grow the interest of a well identified tool using the modelling of information.

Axiomatizing FRBR: An Exercise in the Formal Ontology of Cultural Objects

Allen RENEAR
Yunseon CHOI
Jin Ha LEE
Sara SCHMIDT

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Most current conceptual modeling methods were originally designed to support the development of business-oriented database systems and cannot easily make computationally available many of the features distinctive to cultural objects. Other modeling approaches, such as traditional *conceptual analysis* can complement and extend contemporary conceptual modeling and provide the computing humanist with methods more appropriate for cultural material and humanistic inquiry.

The Humanities and the Problem of Method

Dilthey famously distinguished the methods of the cultural sciences from those of the natural sciences, claiming that the natural sciences seek to explain whereas the sciences of culture seek to understand as well. Although there is no generally accepted account of this distinction, it is still a not uncommon belief that when humanists analyze, explain, and interpret the cultural world, they are, at least in part, using distinctive methods. The question has a long history but it is now especially acute in the practice of humanities computing.

One compromise is to accept the separation and treat computational support as preliminary or ancillary — or, even if constitutive, partial, and the lesser part. We believe that such a resolution will result in missed opportunities to develop intrinsic connections between the methods of managing computational support and traditional methods of advancing humanistic insight.

Conceptual Analysis

The early Socratic dialogues focus on cultural concepts such as justice, piety, courage, beauty, friendship, knowledge, and so on. Socrates asks what these are and attempts to determine what features are significant, sometimes considering hypothetical cases to elicit modal intuitions, sometimes reasoning discursively from general principles. This now familiar style of reasoning may be called "conceptual analysis", or, when formalized with the general principles articulated as axioms, "axiomatic conceptual analysis". Often discussions of cultural objects by humanist scholars can be seen to be some variation of this sort of reasoning or situated within a framework of concepts which could be explicated in this way.

This approach to understanding cultural facts has been widely criticized, from both hermeneutic and positivist quarters; however recent work on the nature of social facts may provide some support. Searle and others have argued that social and cultural facts are established through acts of "collective intentionality" (Searle, 1995). If so then at least part of the nature of that reality would seem to be directly accessible to the participating agents. We cannot investigate galaxies, electrons in this way, because we in no way create them as we do poetry, music, and social institutions. Searle's account is consistent with the approach taken by the phenomenologists of society and culture, such as Reinach and Ingarden, as well as with classical Anglo-American philosophical analysis (Smith, 2003).

An Example: Bibliographic Entities

In the *Functional Requirements for Bibliographic Records* (FRBR) the conceptual modeling is explicit and the conceptual analysis latent. In the text of FRBR we read (IFLA, 1998):

Work: "a distinct intellectual or artistic creation"

Expression: "the intellectual or artistic realization of a work in the form of alphanumeric, musical, or choreographic notation, sound, image, object, movement..." (e.g., a text).

Manifestation: "the physical embodiment of an expression of a work". (e.g., an edition).

Item: "a single exemplar of a manifestation". (e.g., an individual copy of a book)

The novel *Moby Dick*, a work, is realized through various expressions, the different texts of Moby Dick, including different translations. Each one of these expressions may be embodied in a number of different manifestations, such as different editions with different typography. And each of these manifestations in turn may be exemplified in a number of different items, the various individual copies of that edition. Each entity is also assigned a distinctive set of attributes: works have such things as subject and genre; expressions a particular language; manifestations have typeface and type size; and items have condition and location.

Below is the "entity relationship diagram" representing these entities and relationships:

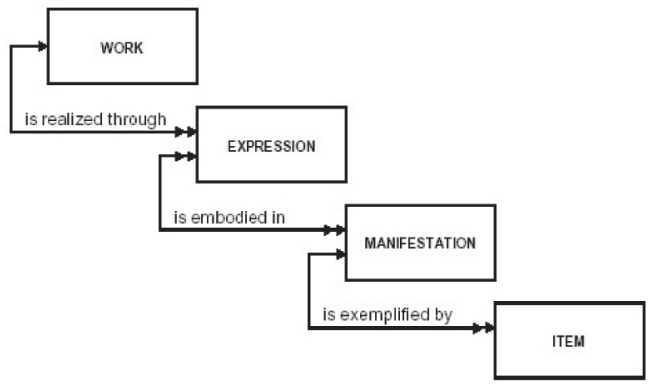


Figure 1: ER Diagram of FRBR Group 1 Entities and Primary Relationships
[diagram from IFLA (1998)]

Entity relationship diagrams are a widely used conceptual modeling technique in the development of information management systems and there are algorithms for converting ER diagrams into robust lower level abstractions, such as normalized relational tables, that can be implemented in database systems. However standard ER diagrams cannot make all aspects of cultural material computationally available. There is no method for saying explicitly under what formal conditions entities are assigned to one entity set or another, for distinguishing entities from relations and attributes, or for identifying necessary or constituent features. Moreover, relationships are understood extensionally, and modal or other intentional assertions, including propositional attitudes and speech acts that are critically important in

the study of society and culture cannot be expressed. (Renear and Choi, 2005).

Extending Conceptual Modeling with Conceptual Analysis

[*Caveat:* In what follows we intend no position on the plausibility of any ontological theory of cultural objects. Our claim is only that there are such positions, that they cannot be easily represented with current conceptual modeling techniques, and that they can be represented with other techniques.]

The text of FRBR provides much information that, despite appearances, is not represented in the FRBR ER diagram. Some of this disparity has been discussed elsewhere (Renear and Choi, forthcoming); here we take up features especially relevant to cultural material.

For example, the FRBR ER diagram does show embodiment, realization, and exemplification relationships, of course, but it does not indicate their particular significance. The "is" of "...is the physical embodiment..." is not the "is" of mere predication. It is a conceptually constitutive "is": we are being told not just a fact about manifestations, but what manifestations (conceptually) *are*. The cascade of definitions suggests this formalization:

work(x) … x is an artistic or intellectual creation

expression(x) =df ($\exists$y)[realizes(x,y) & work(y)]

manifestation(x) =df ($\exists$y)[embodies(x,y) & expression(y)]

item(x) =df ($\exists$y)[exemplifies(x,y) & manifestation(y)]

Now we can see that the concept of work is taken as a quasi-primitive entity, the three characteristic relationships are also each primitive, and essentially involved in the definitions of the entities, and the appearance of interdefinition is made explicit. Because none of this is modeled in the FRBR ER model, that model does not fully represent FRBR's perspective and, moreover, these features will not be reflected in information systems generated from that model and will not be computationally available for analysis.

Bibliographic Platonism

Already we see inferences not entirely trivial, such as the theorem that bibliographic items imply the existence of corresponding (abstract) manifestations, expressions, and works:

P1 item(v) $\supset$ ($\exists x$)($\exists y$)($\exists z$) [manifestation(x) & expression(y) & work(z) & exemplifies(v,x) & embodies(x,y) & realizes(y,x)]

But consider the converse of that conditional:

A1 work(v) $\supset$ ($\exists x$)($\exists y$)($\exists z$) [item(x) & manifestation(y) & expression(z) & exemplifies(v,x) & embodies(x,y) & realizes(y,z)]

A1, a bibliographic analogue of the Aristotelian thesis that only instantiated universals exist does not follow from the definitions.

Represented in this way FRBR now raises a traditional problem for Platonist ontologies of art: if works are abstractions existing independently of their instantiations, then how can they be *created*?

Bibliographic Aristotelianism

An alternative approach could take items as primitive.

work(x) =df ($\exists y$)[IsRealizeBy(x,y) & expression(y)]

expression(x) =df ($\exists$y)[IsEmbodiedBy(x,y) & manifestation(y)]

manifestation(x) =df ($\exists$y)[IsExemplifiedBy(x,y) & item(y)]

item(x) =df a (material) artistic or intellectual creation

This is a "moderate" realism in which **A1** is now a theorem and **P1** no longer one. Here abstract objects cannot exist independently of their physical instantiations, although they do exist (as real objects) when their corresponding items exist. However **P1** will certainly need to be added as an axiom to support our intuition that items do imply works in any case. Or, another approach to the same end is to leave work as primitive, as before, but add **P1** as an axiom. Either might better fits our commonsense intuitions about artistic creation, But we may now have problems characteristic of moderate realism: how to exclude abstract objects which have an intermittent being, going in and out of existence as their instances do — which would be in contradiction to another commonsense intuition: that "a thing cannot have two beginnings in time" (Locke).

A Third Way

In FRBR the notion of a work seems poorly accounted for, tempting further development. Jerrold Levinson defends this definition of musical work:

x is a musical work =df

x is a sound/performance_means-structure-as-indicated-by-S-at-t.

Levinson argues that works are "initiated types" (other examples: the Ford Thunderbird and Lincoln penny) which do not exist until *indication* but once created exist independent of their concrete instances. Our intuitions about artistic creation are now accommodated, but at a cost: a special class of abstract object which, at least arguably, has a beginning in time but never an end, as in Karl Popper's "third world" of cultural objects. Revising the formalization to represent this view is left to the reader as an exercise. It is a little harder than you might think.

International Federation of Library Associations (1998)

Levinson, J (1980). "What a Musical Work Is," *The Journal of Philosophy* 77.

Renear, A. and Choi, Y. (2005) "Trouble Ahead: Propositional Attitudes and Metadata". Proceedings of the 68th Annual Meeting of the American Society for Information Science and Technology (ASIST). Charlotte NC.

Renear, A. and Choi, Y. (forthcoming) "Modeling our Understanding, Understanding Our Models — The Case of Inheritance in FRBR".

Searle, J. (1995). *The Construction of Social Reality*. 1995.

Smith, B. (2003) '"John Searle: From Speech Acts to Social Reality", in B. Smith (ed.) *John Searle*. Cambridge University Press.

Collaborative Scholarship: Rethinking Text Editing on the Digital Platform

Massimo RIVA
Vika ZAFRIN

Brown University Italian Studies

INTRODUCTION

Based at Brown University, the Virtual Humanities Lab [1] is one of twenty-three "models of excellence" in humanities education, supported by the National Endowment for the Humanities for 2004-06. [2] The project is being developed by the Department of Italian Studies in collaboration with Brown's Scholarly Technology Group and with scholars in the U.S. and in Europe.

This paper will report on the achievements of VHL's work during the first two years of its existence as a platform for collaborative humanities research. We will discuss the editing process as we envision it: as a form of interdisciplinary and collaborative knowledge work. We will present issues arising from our experiment with subjective (or "idiosyncratic") text encoding; challenges we face in organizing the work of an international group of collaborators and the procedures for that work; and the process of annotating and indexing large texts collaboratively. Finally, we will hint at VHL's potential applications for pedagogical purposes.

TEXT ENCODING

We have made three Early Modern Italian texts available online: Giovanni Boccaccio's Esposizioni sopra la Comedia di Dante; portions of Giovanni Villani's Cronica Fiorentina; and Conclusiones Nongentae Disputandae by Giovanni Pico della Mirandola. These three texts were selected as representative of different textual typologies (commentary, chronicle and treatise) that solicit different encoding and annotating strategies. The first two are large (around 700 and 200 modern print pages respectively) and heavily semantically encoded.

The third is organized as a textual database meant to provide a flexible platform for annotation. All texts share the technical infrastructure of the VHL.

The encoding was performed along interdisciplinary lines by scholars of Italian literature and history. The Cronica was encoded by two collaborators; the Esposizioni had one principal encoder and several researchers investigating specific issues. All three encoders were asked to annotate without a DTD, using whatever elements they deemed appropriate based on two criteria:

– that the categories elucidated by the encoding are broad enough to produce interesting search results; and

– that, in their estimation, researchers interested in these texts would generally find their encoded aspects interesting as well.

All three encoders received training; further guidance was available upon request. None of the three scholars had had previous semantic encoding experience.

Although the encoding proceeded separately for each text, similarities in what seems most interesting have emerged. Both texts contain encoding of proper names (including people, places, literary works mentioned) and the themes most prevalent in the narratives. These similarities, and the exigencies of the Philologic [3] search engine being built, have prompted us to homogenize encoding across texts, and make it TEI-compliant to the extent possible. It is important to note that this step was taken after the encoding was completed. This afforded our encoders freedom of analytical thought without burdening them with an unfamiliar and very complex set of encoding guidelines.

The encoding process itself presented a challenge on several fronts. As often happens, it took the encoders a while to get used to doing work almost entirely at the computer. Because of the collaborative nature of the project, and because it is good practice in general, we used a versioning system, and the encoders had to deal with the necessity of having an internet connection at least at the beginning and end of each working interval. The two editor-encoders of the Villani text faced particular challenges, working as they were at different institutions, neither of which is Brown. So in addition to juggling the unintuitive (to them) practice of encoding with their other commitments, they faced the need to coordinate their schedules and responsibilities within the editing process. Combined with sometimes unreliable internet access, these circumstances channeled most communication into email and our work weblog.

Blogging, particularly posting incomplete reflections on a work in progress, was initially an obstacle. However, this work has brought two remarkable benefits. First, we have received feedback from people not directly involved in the project. Second, each participant was constantly updated on others' progress. This gave all involved an idea of where VHL as a whole was going, and encouraged discussion at the grant-project level.

ANNOTATING

Built by the Scholarly Technology Group, the annotation engine allows scholars with sufficient access privileges to annotate texts. [Figure 1] Annotations can be visualized, anchored to one or more passages of one or more texts. [Figure 2] A contributor's own annotations can always be modified or deleted by that contributor. At the moment, a feature of the annotation interface is in development that will allow (again, registered) scholars to reply to annotations made by others. We hope that this will foster collaborative thinking and maintain an informal, workshop-like environment for research.

An international group of around thirtyforty-five scholars has agreed to begin annotation of the Pico text. Having completed first-pass encoding of the other two works, we are assembling similar groups for them as well. Invited annotators serve as alpha testers of the search and annotation engines. Depending on the results, we plan to open up the process to the scholarly web community at large. One issue we face is whether to leave annotation open-ended (according to individual scholars' interests and will) or to provide stricter guidelines – a working plan to be followed by all annotators. For now we have opted for an open process: participants will be free to annotate the portions of the text that they prefer. The VHL discussion forum provides a venue were issues arising from the annotating process may be critically addressed.

INDEXING

We have generated indexes of the encoded texts. Merely compiling them took weeks –

automatically generated lists revealed encoding mistakes to be corrected, and highlighted many entries to be researched further. We are not yet confident in the indexes' accuracy, but have neither time nor resources to properly address the issue by ourselves. Here, again, the feedback of the scholarly community will be essential.

We see this as an opportunity to test out the already mentioned discussion forum that completes the VHL toolkit, and to gather potentially interested users for alpha-testing and feedback on the collaboration process itself. A call for participation was disseminated on relevant mailing lists, and mailed directly to relevant academic departments at many North American and European universities. We aim to gather a group of qualified [post]graduate and undergraduate students to help us verify the sizable indexes. At the time of this writing (March 2006), several young scholars have expressed interest in contributing.

CONCLUSIONS AND A LOOK TO THE FUTURE

As a final step during the present grant period, we are organizing the existing toolset (search and annotation engines, indexes, weblog and discussion forum) under an umbrella category of the Virtual Seminar Room, which will serve as the venue where editing practices will be consistently linked to pedagogical activities. This move is prompted in part by the success of the Decameron Web's Pedagogy section, which continues to receive positive feedback from teachers of Boccaccio all over the world.

It is too early to state definitively how the VHL will be used by humanists. Based on our prior experience with the Decameron Web [4] and the Pico Project [5], however, we are cautiously optimistic. It is true that work performed entirely online, and collaboration as a mode of research, have been slow to catch on in the humanities. Recent publications and tool developments point to a desire on the part of humanist academics to have spaces akin to science labs, where they can mingle and talk informally about their research. [6] Such labs are difficult and impractical to set up in physical space. So we have created a place online where scholars may interactively edit and annotate texts, and develop pedagogical modules for their individual purpose. With user feedback, we hope to make the VHL attractive enough to humanities scholars that they'll be convinced to come play with us, even if the modes of interaction may be unusual or confusing at first.

The past two years have resulted in a long wishlist of features to implement in the future, given time and resources:

- addition of automatic lemmatizers and other pedagogical parsing and mapping tools, aimed at the various textual typologies of VHL content;
- hosting and inclusion of texts uploaded by users;
- possibility of using the editing process as part of a seminar-like pedagogical experience;
- possibility of adding images as a consistent part of the editing/illustrating process;
- tools for transcription of manuscripts and incunabula; and others.

The future of the humanities is shaping up to be both online and collaborative. The question is not whether humanists will work together, but where they will do so, and what forms their knowledge work will take in the public research arena provided by the web. The VHL is one practical step towards answering this question.

[1] http://www.brown.edu/Departments/Italian_Studies/vhl/

[2] National Endowment for the Humanities. (2004) "NEH Grants Support Models of Excellence in Humanities Education: $3.8 million awarded to 23 projects to create new humanities resources and develop new courses." http://www.neh.gov/news/archive/20040419.html

[3] http://philologic.uchicago.edu/

[4] http://www.brown.edu/decameron/

[5] http://dev.stg.brown.edu/projects/pico/

[6] Gina Hiatt's 2005 article in Inside Higher Ed (http://insidehighered.com/views/2005/10/26/hiatt) is a good example of discussion on the topic. "What I am advocating," she writes, "is injecting into the humanities department some of the freewheeling

dialogue found in the halls outside the conference presentation or in some of the better scholarly blogs." Tool makers have heeded the siren song of collaboration as well; resources such as TAPoR (http://tapor.mcmaster.ca/) and the Virtual Lightbox (http://www.mith2.umd.edu/products/lightbox/) attest to this.

Use of Computing Tools for Organizing and Accessing Theory and Art Criticism Information: the TTC - ATENEA Project

Nuria RODRÍGUEZ ORTEGA
U. Málaga (Spain)

Alejandro BÍA
Juan MALONDA
U. Miguel Hernández

INTRODUCTION. BACKGROUND AND INITIAL GOALS

The main purpose of this article is to introduce and describe a new digital resource, the Tesauro Terminológico-Conceptual (TTC), designed to perform research online, and the Atenea Corpus. The structure of the TTC thesaurus is conceptually innovative, and it has been specifically build to assist the user in the epistemological, terminological, textual and theoretical aspects of art research.

The TTC is a development of the research done by Nuria Rodriguez for her doctoral dissertation [1]. The goal was to find a satisfactory answer to the terminological/ conceptual problems that hinder research in the history, theory and criticism of art, and the TTC is the solution proposed(1). This project is closely related to the Getty Research Institute (Los Angeles, USA), and particularly to its Vocabulary Program, and the Art & Architecture Thesaurus (AAT).(2) The TTC is also linked to the Spanish version of the AAT, translated and managed by the "Centro de Documentación de Bienes Patrimoniales" based in Chile.(3)

The TTC can be broadly defined as a knowledge tool for art specialists and other users, that helps to obtain a compilation of described, analyzed, classified and linked terms and concepts, primarily those having to do with theory, criticism and aesthetics. In order to build this type of tool it is essential to perform an interpretative analysis of these concepts and terms due to the ambiguity that

define them. After that, these terms and concepts must be incorporated into the structure of the TTC with all the relevant information associated to them. An important outcome of this process is the development of a digitalized set of artistic texts: the Atenea Corpus, since one of the epistemological and methodological basis of this project is the study of the terms in their textual contexts.

Therefore, this project creates a virtual net made up of two complementary components: a text database, and the terminological conceptual thesaurus (TTC). The textual database consists of all the digitalized texts that supply the terms and concepts recorded by the thesaurus [see figure 1].

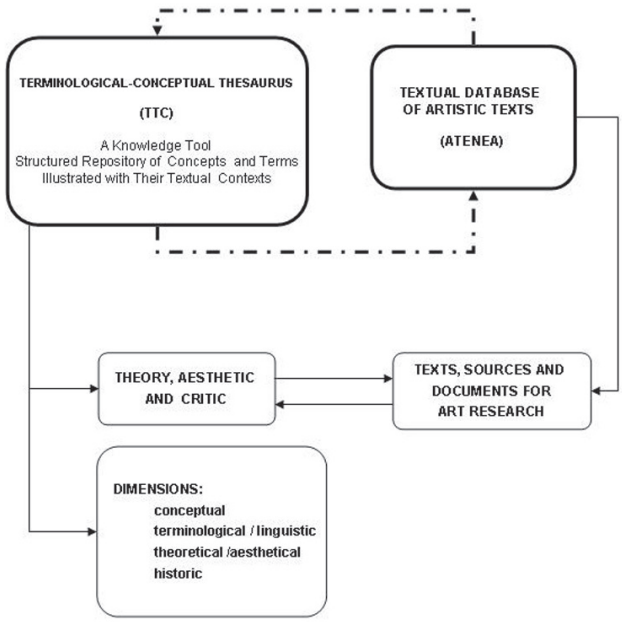


Fig. 1.: the TTC and ATENEA resources.

INITIAL GOALS

Why to create a new type of thesaurus? The initial context in which the project began substantially determined the conception and the characteristics of what would become the TTC. For further information on this type of thesaurus see also [2, 3, 4 and 5].

1. Terminological/conceptual problems: ambiguity and interpretative inaccuracy.

The main purpose of this project was to provide a satisfactory answer to the terminological/conceptual problems that encumber the task of conducting research in the history of art. Some of the problems most frequently found in the field of art history are terminological and discursive ambiguity, vagueness, and inaccuracy, particularly when dealing with theory and criticism. This situation derives from the high degree of semantic density typical of artistic terminology, as well as from frequent semantic alterations and shifts in meaning that terms undergo as a result of multiple reinterpretations and new uses. Thus, studying the semantic properties of terms and their diverse domains of meaning became a priority from the beginning. This approach required identifying and organizing the various meanings of each term and the semantic scope of every concept by comparing the variations found in the different discourses.

None of the available thesauri was capable of performing these tasks: 1. Conventional thesauri do not take into account the theoretical and discursive contexts in which concepts appear; 2. The brief definition (or scope note) attached to each descriptor in conventional thesauri barely satisfies the needs of specialists in theory or criticism. The scope note has a very specific function, which is to limit the meaning of the descriptor to the framework of the thesaurus itself. As a result, theoretical and critical analyses are irrelevant in defining and describing concepts. Lacking these sorts of analyses, the complexity inherent to artistic concepts, with all their nuances and variations, ends up either simplified or ignored.

2. Linguistic and expressive richness.

Besides categorizing the semantic component of the terminology and solving its ambiguities, a further objective was to compile and analyze the rich variety of verbal and expressive resources that characterizes the field of art history. In order to reach this goal it was necessary to collect, describe and classify all the expressive resources used by specialists in their writings about art, whether the lexical expressions were specific, or not (such as metaphoric or metonymic terms, literary or rhetorical resources). The criteria applied to the compilation of expressive resources went beyond the standards of conventional thesauri, for they only register terms identified as belonging to a particular field of study.

3. Textual and discursive approach.

The essential notion underlining the project is that

natural languages have to be considered in the context of their use, as they appear in a particular text; with respect to terms, they are regarded as functional entities with a contextually determined role and meaning. Consequently, the identification and analysis of terms and concepts had to be done at the discursive level, paying especial attention to the way in which authors and specialists used and defined them. Since these terminological and conceptual uses needed to be examined within their textual context, the thesaurus had to include the relevant discursive fragments. Again, conventional thesauri could not offer this type of information because, even though they provide the bibliographical sources for the terms, they do not include the textual citations themselves.In view of the shortcomings of conventional thesauri, it became clear that only the development of a new thesaurus prototype would achieve the desired results.

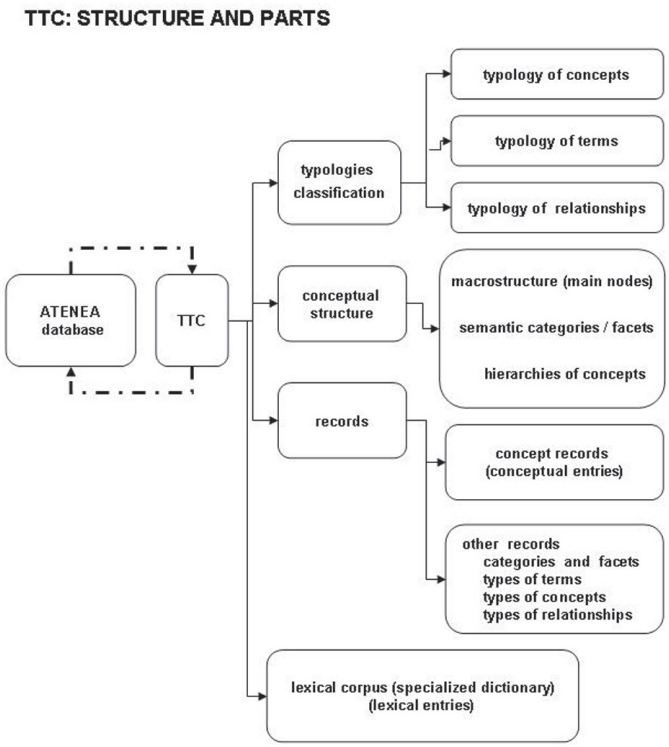


Fig. 2: structure and parts of the TTC.

It follows that the TTC simultaneously develops in three dimensions:

- Conceptual and Epistemological Dimension: comprises a collection of artistic concepts described, classified into a conceptual structure and linked among them.
- Linguistic and Terminological Dimension: consists of every term and verbal expression that acquires artistic meaning in the discourses being analyzed; all of them are described, classified according to terminological typologies, and assigned to the concepts they denote.
- Textual and Discursive Dimension: contains the texts that are the sources for the concepts and terms registered in the TTC.

Thus, the structure of the TTC, as we will explain with detail in the presentation, can be simply described as the way in which these dimensions are organized [see figure 2].

THE TEXTUAL DATABASE

The database comprises the set of Spanish artistic texts used to extract all the concepts and terms registered in the TTC. This is why the textual database and the TTC can be described as two complementary tools that are virtually linked. Once there is a suitable computer application to implement the project, it will be possible to go from the textual fragments found in the TTC to the full textual sources; likewise, it will be possible to access the TTC from the textual database to obtain detailed information about the terms and concepts that appear in the texts.

In addition to its complementary role, the textual corpus has also been designed to function as an independent tool that can be used by any specialist or researcher interested in aesthetics, theory, or art criticism. Indeed, a key goal of this project is to turn the textual corpus ATENEA into an exhaustive repository of pictorial and artistic texts that could serve as a basic reference tool for theoretical and critical studies, and, more generally, for research dealing with texts, sources and documents. This textual corpus represents an important contribution to the study of Spanish art history and theory since there are no other systematically digitalized texts in this field.

The encoding of the texts was done at the Miguel Hernández University(4), following the TEI standards(5). The markup was performed using the Spanish set of TEI tags, allowing for automatic conversion to standard English-based TEI tags when necessary [4]. We considered that marking Spanish texts with Spanish mnemonics is more coherent than using English based marks, and

improve their understanding and usability by Spanish speaking scholars [5].

For Web management and operation we used the PhiloLogic System.(6) This system was developed by the ARTFL project in collaboration with the University of Chicago Library's Electronic Text Services(7) for the purpose of running texts compiled on digital corpuses.

The textual database offers a diversity of search options that are redundant in terms of their versatility and polyvalence [see figure 3]. Depending on their interests, users will be able to:

- Access a digitized copy of the complete text or texts of the author or authors being examined.
- Find a listing of all the data related to a specific term, such as how it is used in a particular treatise, the frequency with which it appear, the various meanings it assumes, or even compare different treatises.
- Get information about the sources or documents themselves. To this effect the textual corpus is being implemented with an additional database containing all the information related to the digitalized texts in specially structured and codified fields.

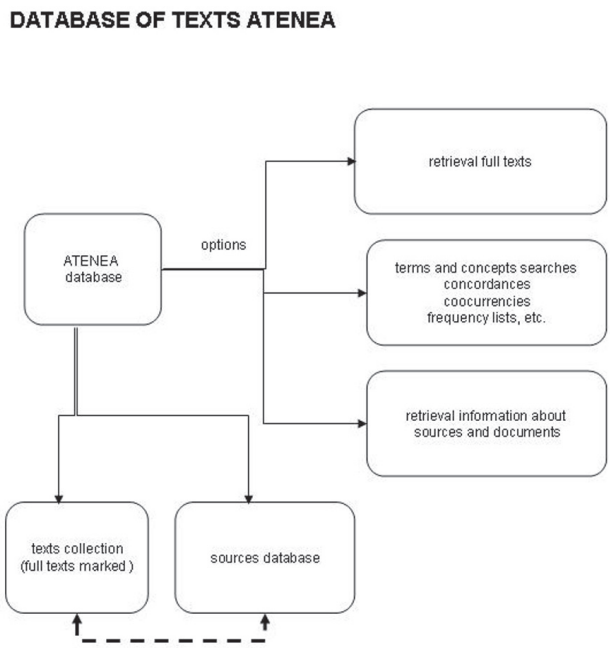


Fig. 3: uses and possibilities of ATENEA database.

USES AND POSSIBILITIES OF THE TTC

The TTC has been conceived and developed as a research tool whose main emphasis is the terminological, conceptual and theoretical/critical analysis of artistic texts. However, the potential uses of the TTC are not limited to textual analyses for it has the capability of becoming a polyvalent tool applicable to a variety fields to the study of art in the Web.

In general terms, there are four main potential applications that deserve to be emphasized:

1. Terminological and conceptual research

2. Tasks related to documentation and indexing of visual resources and textual material

3. In connection with this last point, retrieval of artistic information mounted on the Web.

4. Finally, the structure of TTC can also work as an abstract and standardized model that could be used by other specialists or other projects. These "other" TTCs could be linked as microthesauri.

FOOTNOTES

(1) The development and management of this type of thesaurus was an integral part of a project called *Desarrollo y gestión de recursos cibernéticos para la sistematización , producción y difusión del conocimiento artístico,* based at the University of Málaga, under the supervision of Nuria Rodriguez. Recently, this project, newly called *Desarrollo de un tesauro terminológico-conceptual (TTC) sobre los discursos teórico-artísticos españoles de la Edad Moderna*, *complementado con un corpus de textos informatizado (ATENEA)*, has got a grant from the Ministerio de Educación y Ciencia of Spain, which will support its development during the next three years.

(2) http://www.getty.edu/research/conducting_research/vocabulary.aat

(3 http://www.aatespanol.cl

(4) The markup for this project was performed by Juan Malonda and Alejandro Bia.

(5) Text Encoding Initiative

(6) http://www.lib.uchicago.edu/efts/ARTFL/philologic/index.html

(7) http://www.lib.uchicago.edu/e/ets/

References

[1] **Rodríguez Ortega, N**. (2003). *Tesauro terminológico-conceptual de los discursos teóricos sobre la pintura (primer tercio del siglo XVII)*. Málaga: published on CD-ROM by the Servicio de Publicaciones de la Universidad de Málaga.

[2] **Rodríguez Ortega, N**. (1999). "Un proyecto de estudio en la historia del arte: presentación y descripción de un modelo de tesauro pictórico-artístico (I)". *Boletín de Arte,* 20. pp. 395-421.

[3] **Rodríguez Ortega, N.** (2004). "Una 'realidad' para la investigación en la Historia del Arte: tesauro terminológico-conceptual de los discursos teórico-críticos y estéticos (II)". *Boletín de Arte*, 24. pp. 71-101.

[4] **Rodríguez Ortega, N**. "El conocimiento artístico en Red. Desarrollo de herramientas cibernéticas para la investigación en Historia del Arte: tesauro terminológico-conceptual (TTC) y corpus digitalizado de los discursos teórico-críticos españoles". *Actas del XV Congreso Nacional de Historia del Arte. Modelos, intercambios y recepción artística (De las rutas marítimas a la navegación en red)*. Palma de Mallorca, 20-23 de octubre de 2004 (en preparación).

[5] **Rodríguez Ortega, N**. (2006). *Facultades y maneras en los discursos de Francisco Pacheco y Vicente Carducho. Tesauro terminológico-conceptual*. Málaga: Universidad de Málaga, Academia de Bellas Artes de San Telmo, Colegio Oficial de Peritos e Ingenieros Técnicos de Málaga.

[6] **Bía, A., Malonda, J. and Gómez, J**. (2005). "Automating Multilingual Metadata Vocabularios". *DC-2005: Vocabularies in Practice*. Madrid, Carlos III University, 12-15 September 2005.

[7] **Bía, A. and Sánchez-Quero, M.** (2004). "The Future of Markup is Multilingual". *ACH/ALLC 2004: The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*. Göteborg University, Sweden, 11-16 June 2004, pp. 15-18.

The Ghost of the Printed Page: New Media Poetry and Its Ancestry

Jennifer ROWE

University Of Maryland, Department of English

In 1996, the Visible Language journal published, according to Eduardo Kac, "the first international anthology to document a radically new poetry, one that is impoosible to present directly in books and that challenges even the innovations of recent and contemporary experimental poetics" (98). The enthusiasm for the "radical newness" of new media poetry with which Kac introduces the anthology has been realized in part over the last ten years by a veritable explosion of new media poetry available online and in CD/DVD-ROM format.

The interest of new media poets in semantic experimentation as well as with investigations of materiality have lead most critics to locate the roots of this genre in the tradition of the many avant-garde movements and experimental poetries of the twentieth-century, including surrealism, cubism, Dadism, forms of visual poetry, the L=A=N=G=U=A=G=E movement and futurism, just to new a few. Not surprisingly, this critical approach has tended to favor artists and those works, like Kac's "Insect.Desperto" and Melo e Castro's "The Cryptic Eye," that most noticeably demonstrate certain tendancies of the avant-garde such as multi or non-linearity, non-narrative or non-grammatical strings of words, and the priveleging, to use Roberto Simanowski's term of the "optic gesture" of the word over its "semantic meaning" (7).

By contrast, works that fall under the purview of new media poetry by nature of their digital rendering and their use of multi-media software, but that, nonetheless, bear resemblances to more traditional twentieth-century poetries are often dismissed from consideration. This presentation will begin by exploring this critical resistance to digital work that bears resemblance to traditional print-based poetry; a resistance, I will argue, that belies the great number of new media poems that incorporate formal techniques and conventions characteristic of non-experimental poetry--in particular, techniques and conventions typical of the free verse lyric style that characterized much of the poetry of the twentieth century.

By glancing at any of the online databases for new media poetry, we can see poets utilizing conventions such as iambic meter, stanzaic organization, alliteration, white space, and enjambment, among others, to organize and present the poetic line to the reader. For example, in Farah Marklevit's "How They Sleep" the author uses a combination of end-stopped and enjamed lines to juxtapose what can be said to be the internal movement of the verse with the scrolling motion of the text on screen. The stop/start motion of the poem underscores the story being told--a romantic struggle between a husband and wife whose heads are "tuned to different pitches / like glasses of water." As we can see in this example, the use of the rolling text serves not as an alternative to enjambment and punctuation, but rather as a vehicle by which these formal techniques are more fully realized by their incorporation into the multimedia presentation.

The above is just one example of how attention to traditional formal poetic conventions can enhance our understanding of the aesthetic strategies of new media work; there are many others. It will be the contention of this presentation that, contrary to what most of the existing scholarship on new media poetry would suggest, the influence of traditional poetic forms and techniques is pervasive in the genre of electronic literature. Rather than rendering these formal strategies obsolete, moreover, the electronic environment offers poets the opportunity to render these strategies in ways that cannot be replicated on the printed page. In addition, I will attempt to show how situating new media poetry solely within the heritage of experimental poetry movements can not fully account for the aesthetic qualities of a great number of works.

Rather than attempt an inclusive survey of these works, my presentation will focus specifically on the web-based work of Thomas Swiss, a poet who began, and continues, to write in print format and has recently forayed into the genre of new media poetry. I will focus on those poems that both explore the ramifications of new writing technologies as well as lay bare their debt to more traditional print-based poetry. It will be my argument that it is precisely by utilizing formal conventions typical

of the free verse lyric and by preserving the visual style of the printed poem that Swiss interrogates the ways in which our pre-established patterns of reading are affected by the migration of the poem from page to screen.

Non-traditional Authorship Attribution : The Contribution of Forensic Linguistics

Scientific testimony based on linguistic stylistics has determined the outcome of many civil and criminal cases involving questions of authorship

McMenamin (48)

Joseph RUDMAN

Department of English,
Carnegie Mellon University, USA.

INTRODUCTION

Black's Law Dictionary (648) defines forensic linguistics as:A technique concerned with in-depth evaluation of linguistic characteristics of text, including grammar, syntax, spelling, vocabulary and phraseology, which is accomplished through a comparison of textual material of known and unknown authorship, in an attempt to disclose idiosyncracies peculiar to authorship to determine whether the authors could be identical.

At the 2002 ALLC/ACH Conference in Tuebingen, Laszlo Hunyadi et al. discussed some of the contributions that humanities computing makes to forensic linguistics. In this paper I point out the many contributions that forensic linguistics has and is making to the larger field of non-traditional authorship attribution -- contributions that are unknown or largely ignored by most non-forensic practitioners of non-traditional authorship attribution (this statement is based on the lack of references to the wealth of studies in forensic linguistics -- a quick glance at the bibliography will show some exceptions):

1) Immediacy
2) Techniques
3) Scientific Validity
4) Gatekeeping

5) Levels of Proof
6) Rules of evidence
7) Organization

The paper goes on to propose closer formal ties between the ALLC/ACH and the IAFL. In each of the following sections, there is an emphasis on how the forensic techniques should be employed by the non-forensic practitioner and on the contributions of humanities computing to the field.

IMMEDIACY

This section discusses how immediacy forces a more careful, more restrictive methodology on the forensic practitioner (versus the non-forensic):

> Forensic linguistics is a sub-set of authorship attribution that is much more immediate and in many ways demands a "correct" attribution. Forensic linguistics often deals in criminal guilt or innocence -- with serious ramifications -- even life or death!

TECHNIQUES

This section discusses:

1) The necessity of employing corpus linguistic techniques.

   a) Looking at style markers as deviations from the norm.

   b) Looking at grammatical, stylistic, spelling, punctuation, and orthographical errors as style markers.

   c) The limiting of one control group to "suspects."

2) The need for non-forensic practitioners to acquire the skills necessary to navigate bibliographic research in forensics (primary and secondary resources in case law -- and the many commentaries).

SCIENTIFIC VALIDITY

This section discusses:

1) The concept that all scientific methods should be brought to bear on an authorship problem -- e.g. handwriting analysis, paper analysis, type font analysis.

2) The strict definition of expert witness -- and the role they are allowed to play -- e.g. guides to help the jury interpret the facts.

3) How methodology must be "generally accepted by the community of scholars" to be allowed.

GATEKEEPING

This section discusses:

1) The role of the courts (of various countries) as gatekeepers.

2) The role of the International Association of Forensic Linguists IIAFL as gatekeeper and certifier of gatekeepers.

This does not mean that there are not flaws in the system -- e.g. allowing the Morton CUSUM method as a valid technique even after its debunking on live TV.

LEVELS OF PROOF

This section discusses:

> How the "answer" is presented - while non-forensic practitioners for the most part present "probabilities", forensic linguistics presents a "preponderance of evidence" concept and one of being "beyond a reasonable doubt."

RULES OF EVIDENCE

This section discusses:

> The intricacies of the rules of evidence and how these rules can give direction to non-forensic attribution studies. Rules of evidence are not universal -- different countries have different rules -- in the United States, different states have different les. A set of rules distilled from all those available is advocated or the non-forensic practitioner.

a) Dauber

b) Post-Dauber

ORGANIZATION

This section discusses:

1) The IAFL, a "professional" organization with the requirement that its full members show evidence of "linguistic qualifications."

a) The IAFL's journal -- Speech, Language and the Law (Formerly - Forensic Linguistics) Having a

paper published here -- the implicit nihil obstat of the IAFL -- gives added weight to the credentials of the practitioner. Among other important journals are, Expert Evidence, Forensic Science International, and Journal of Forensic Document Examination.

b) The IAFL's conference

The IAFL holds a biennial conference. The last one was in July, 2005 at Cardiff University, UK. Seven of the presented papers are of interest (and importance) to non-forensic practitioners -- e.g. Sanchez et al.'s "Intra and Inter-author Comparisons: The Case of Function Words: Are Function Words Really Functional in Stylometric Studies of Authorship Attribution."

c) There are some members (formal and contributing) of ALLC and ACH that are also members of the IAFL but their work published in non-forensic journals is quite different.

CONCLUSION

Forensic linguistis is not a "perfect" discipline. One unfortunate aspect of forensic linguistics is the adversarial role in presenting evidence -- many forensic linguistic presentations, while not necessarily fraudulent or even unethical are not in the best interests of practitioners who want to present the "whole truth." Another unfortunate side-effect of the judicial system on the complete reporting of authorship studies is the all too common practice of "sealing" court records when a settlement is reached outside of the courtroom. I have seen some of these sealed records and only hope that the techniques will be duplicated and published elsewhere. I do not want to give the impression that non-forensic attribution is a poor cousin with nothing to offer -- the many disciplines that form the bulk of the field (e.g. computer science, stylistics, statistics) are the "core" -- but this is for another time.

References

Aitken, C. G. G. (1993). "Conference Report: Statistics and the Law." Journal of the Royal Statistical Society -- A Part 2. 156: 301-304.

Aked, J P., et al. (1999). "Approaches to the Scientific Attribution of Authorship." In Profiling in Policy and Practice. Eds. David Canter and Laurence Alison. Aldershot: Ashgate. pp. 157-187.

Barnes, D W. (1983) Statistics as Proof: Fundamentals of Quantitative Evidence. Boston: Little, Brown and Company.

Belkin, R., and Y. Korukhov (1986). Fundamentals of Criminalistics. Moscow: Progress Publishers. (Translated from the Russian by Joseph Shapiro.) (See Chapter 5, "Criminalistic Study of Documents.")

Bailey, R W. (1979). "Authorship Attribution in a Forensic Setting." In Advances in Computer-Aided Literary and Linguistic Research: Proceedings of the Fifth International Symposium on Computers in Literary and Linguistic Research. Eds D.E. Ager, F.E. Knowles, and Joan Smith. Birmingham: John Goodman. pp. 1-15.

Black, B., et al. (1997). "The Law of Expert Testimony -- A Post-Daubert Analysis." In Expert Evidence: A Practitioner's Guide to Law, Science, and the FJC Manual. Eds. Bert Black and Patrick W. Lee. St. Paul, Minn.:West Group. pp. 9-71.

Brautbar, N. (1999). "Scientific Evidence." In Ethics in Forensic Science and Medicine. Ed. Melvin A. Shiffman. Springfield, Illinois: Charles C. Thomas. pp. 92-121.

Campbell, D. (1992). "Writing's on the Wall." The Guardian Wednesday 7 October.

Canter, D. (1992). "An Evaluation of the 'Cusum' Stylistic Analysis of Confessions." Expert Evidence 1.3: 93-99.

Carpenter, R H. (1990). "The Statistical Profile of Language Behavior with Machiavellian Intent or While Experiencing Caution and Avoiding Self-Incrimination." In The Language Scientist as Expert in the Legal Setting: Issues in Forensic Linguistics. Eds. Robert W. Rieber and William A. Stewart. New York: The New York Academy of Sciences. pp. 5-17.

Chaski, C E. (2001). "Empirical Evaluations of Language-Based Author Identification Techniques." Forensic Linguistics 8.1: 1-65.

Chaski, C E. (1997).“Who Wrote it?: Steps Toward a Science of Authorship Identification.” National Institute of Justice Journal 233: 15-22.

Coulthard, M. (1992). “Forensic Discourse Analysis.” In Advances in Spoken Discourse Analysis. Ed. Malcolm Coulthard. London: Routledge. pp. 242-258.

Drommel, R H., and U W. Lohr. (1990). “Text Attribution in a Forensic Setting.” ALLC-ACH 90 Conference. University of Siegen. Siegen, Germany. 4-9 June.

Eagleson, R D. (1989). “Linguist for the Prosecution.” In Words and Wordsmiths: A Volume for H. L. Rogers. Eds. Geraldine Barnes et al. Sydney: The University of Sydney. pp. 22-31.

Finegan, E. (1990). “Variation in Linguists’ Analyses of Author Identification.” American Speech 65.4: 334-340.

Fitzgerald, J R. (2002). “The Unabom Investigation: A Methodological and Experimental Study from a Forensic Linguistic Perspective.” Preprint from the author. July.

Frederico, D R., and R A. Weiner. (2000). “Owning Daubert: Challenging Expert Testimony as a Defense Strategy.” For The Defense 42.11: 12-13, 47-48.

Frye v. United States, 54 App. D.C. 46, 293 F. 1013 (1923).

Gallagher, M C. (1983). “Linguistic Evidence: Making a Case for Admissibility.” Legal Times (Washington, D.C.) July 4: 14, 19-20.

Grant, T., and K. Baker. (2001). “Identifying Reliable, Valid Markers of Authorship: A Response to Chaski.” Forensic Linguistics 8.1: 66-79.

Hardcastle, R A. (1993). “Forensic Linguistics: An Assessment of the CUSUM Method for the Determination of Authorship.” Journal of the Forensic Science Society 33.2: 95-106.

Holmes, D I., and F J. Tweedie. (1995): “Forensic Stylometry: A Review of the Cusum Controversy.” Revue Informatique et Statistique dans les Sciences Humaines 31: 19-47.

Hunyadi, L., et al. (2002). “Forensic Linguistics: The Contribution of Humanities Computing.” Paper delivered at ALLC/ACH 2002. University of Tuebingen 24-28 July.

Iancu, C A., and P W. Steitz. (1997). “Guide to Statistics.” In Expert Evidence: A Practitioner’s Guide to Law, Science, and the FJC Manual. Eds. Bert Black and Patrick W. Lee. St. Paul, Minn.: West Group. pp. 267-318.

Klein, M S. (1997). “Empowering the Gatekeeper in the Post-Daubert Regime: Court-Appointed Experts and Special Masters. In Expert Evidence: A Practitioner’s Guide to Law, Science, and the FJC Manual. Eds. Bert Black and Patrick W. Lee. St. Paul, Minn.: West Group. pp. 425-450.

Kniffka, H. (1996). “On Forensic Linguistic ‘Differential Diagnosis.’” In Recent Developments in Forensic Linguistics. Ed. Hannes Kniffka (In cooperation with Susan Blackwell and Malcolm Coulthard). Frankfurt am Main: Peter Lang. pp. 75-121.

Kredens, K. (2004). “Investigating Idiolect Performance: Towards a Methodology of Forensic Authorship Attribution.” [Description of the Research Project.] Copy provided by author, 24 February.

Loue, S. (1999). “Junk Science and Frivolous Claims.” In Ethics in Forensic Science and Medicine}. Ed. Melvin A. Shiffman. Springfield, Illinois: Charles C. Thomas. pp. 122-128.

Matthews, R. (1993). “Linguistics on Trail.” New Scientist 139.1887: 12-13.

Matthews, R. (1993). “Harsh Words for Verbal Fingerprints.”Sunday Telegraph 4 July.

McMenamin, G R. (2002). Forensic Linguistics: Advances in Forensic Stylistics. Boca Raton: CRC Press.

McMenamin, G R. (2001). “Style Markers in Authorship Studies.” Forensic Linguistics 8.2: 93-97.

McMenamin, G R. (1981). “Forensic Stylistics.” [Chapter 39D (Vol. 4), (Rel. 17-6/94 Pub. 313)] In Forensic Sciences. Gen. Ed. Cyril H. Wecht. New York: Matthew Bender & Co., Inc. [5 Vol. (loose-leaf) -- A continually updated series.]

McMenamin, G R. (1993). Forensic Stylistics New York: Elsevier, 1993. (Reprinted from Forensic

Science International 58.)

Menicucci, J D. (1978). "Stylistics Evidence in the Trial of Patricia Hearst." Arizona State Law Journal 1977.2: 387-410.

Miron, M S. (1990). "Psycholinguistics in the Courtroom." In The Language Scientist as Expert in the Legal Setting: Issues in Forensic Linguistics. Eds. Robert W. Rieber and William A. Stewart. New York: The New York Academy of Sciences. pp. 55-64.

Morgan, B. (1991). "Authorship Test Used to Detect Faked Evidence."The Times Higher Education Supplement (London)} 9 August.

Morton, A Q. (1991). "The Scientific Testing of Utterances: Cumulative Sum Analysis." Journal of the Law Society of Scotland September: 357-359.

Mullin, J. (1991). "Confession Test Clears Prisoner." The Guardian Thursday 11 July.

Musilova, V. (1993). "Forensic Linguistic Examination of Anonymous Communications." Journal of Forensic Document Examination 6: 1-13.

Niblett, B., and J. Boreham. (1976). "Cluster Analysis in Court." Criminal Law Review March: 178-180.

Nickell, J. (1996). Detecting Forgery: Forensic Investigation of Documents. Lexington: The University Press of Kentucky.

Osborn, A S. (1975). The Problem of Proof: Especially as Exemplified in Disputed Document Trials (2nd Edition), 1926. A Facsimile Edition. Chicago: Nelson-Hall Inc.

Osborn, A S. (1978). Questioned Documents (2nd Edition), 1929. A Facsimile Reproduction. Chicago: Nelson-Hall Co.

Sanford, A J., J P. Aked, L M. Moxey, and J Mulin. (1994). "A Critical Examination of Assumptions Underlying the Cusum Technique of Forensic Linguistics." Forensic Linguistics 1,2: 151-167.

Smith, M W A. (1989). "Forensic Stylometry: A Theoretical Basis for Further Developments of Practical Methods." Journal of the Forensic Science Society 29: 15-33.

Svartvik, J. (1968). The Evans Statements: A Case for Forensic Linguistics. Stockholm: Almqvist & Wiksell.

Totty, R N., R A. Hardcastle, and J. Pearson. (1987). "Forensic Linguistics: The Determination of Authorship from Habits of Style." Journal of the Forensic Science Society 27: 13-28.

United States v. Hearst, 412 F. Supp. 893 (1976).

Winter, E. (1996). "The Statistics of Analysing Very Short Texts in a Criminal Context." In Recent Developments in Forensic Linguistics. Ed. Hannes Kniffka (In cooperation with Susan Blackwell and Malcolm Coulthard). Frankfurt am Main: Peter Lang. pp. 141-179.

Proposing an Affordance Strength Model to Study New Interface Tools

Stan RUECKER

Humanities Computing in English and Film Studies, University of Alberta, Canada

This paper suggests a new approach to the study of potential interface tools, by examining not the tools themselves, but instead a set of factors that contribute to the possible benefit that might be provided by the tools. Since the proposed "affordance strength model" does not require a working version of the tool to study, it can therefore be applied at several points, beginning even very early in the research cycle, at the initial concept stage. Many standard usability instruments, such as GLOBAL, include questions that cover different aspects of the user's perception of the tools, but require a working prototype. Other usability protocols exist for studying systems at an early design stage, such as TAM (Morris and Dillon 1997). However, these do not include a complete range of the factors in the proposed "affordance strength model." This proposed model can also be used at later stages, both at the point where prototypes have been created, and later still, once working versions are in production. Researchers can also begin to compare the affordance strength of different kinds of software tools.

An affordance is an opportunity for action (Gibson 1979). For computer interface designers attempting to create new software tools—that may in some cases offer new opportunities for action—a perennial problem exists concerning how best to study an affordance that was not previously available. Comparisons against previous interfaces with different affordances tend toward category error (comparing apples to oranges), and comparisons against interfaces with similar affordances but different designs tend to be studies of design rather than of opportunities for action.

Given the need to specify the significant relational factors that characterize the strength of an affordance, it is possible to distinguish eight factors that together represent the relational aspects of the object, the perceiver, and the dynamics of the context. These factors together can be used to create a vector space that defines the relational aspects of affordance strength in an operational way. Although these eight factors are not the only possible candidate factors, it is possible to explain how they work together to create a relatively strong picture of the total affordance strength:

affordance strength = (tacit capacity, situated potential, awareness, motivation, ability, preference, contextual support, agential support).

Tacit Capacity

The first necessary factor is the tacit capacity of the object to provide the affordance in situations of the kind being studied. For example, if a given adult wishes to keep dry while walking two blocks in the rain, the unfactored affordance of the object is the twin capacity to be carried while walking and, simultaneously, keep someone dry. In this case, the tacit capacity of the umbrella in situations where a person needs to walk two blocks in the rain while staying dry would be very high, while the tacit capacity of, for example, a wrench, would be zero. The wrench has an excellent tacit capacity for other types of actions. In fact, because it is a specialized tool (like the umbrella), it has a primary affordance. But for the work at hand it is useless.

Situated Potential

The second necessary relational factor is the situated potential of the object, not generally in circumstances of the kind under investigation, but in one particular situation at one particular time. It is all very well for the person about to walk in the rain to realize that an umbrella has an excellent tacit capacity for keeping a person dry, when at the point of setting out there is no umbrella available, or the umbrella that is available is torn.

These two factors – tacit capacity and situated potential – are relational attributes where the attention of the researcher is directed toward the object or environment and its relevant affordances for action. There are other factors that treat the relational aspects of the agent, where the researcher's attention is directed at what have been called the perceiver's effectivities (Turvey and Shaw 1979).

Awareness

The first of these factors is awareness. For the person about to walk in the rain, a perfectly good umbrella might be sitting to hand, but if the person is distracted or confused or in a rush, the umbrella might not be perceived, and for all of its high tacit capacity and situated potential, the umbrella still stays dry while the person gets wet.

Motivation

The second factor is motivation. If the person in question wants to walk in the rain and would prefer not to get wet but does not really mind it all that much, that person's tendency to seek and adopt an available affordance is significantly reduced in comparison with the person who hates getting wet, has just had a cold, and is wearing clothes that will be damaged by the rain. The former person may casually take up an available umbrella if one were available, since the tacit capacity and situated potential are high enough that the action has an appropriately low resource load. If only a newspaper is available, the lower tacit capacity might be such that the person would prefer to simply get rained on. For the latter person, it is likely that the high motivation and absence of an umbrella would lead to extremes of behavior such as deciding not to walk but take a taxi instead, or perhaps going back into the building to see if an umbrella could be found somewhere.

Like many of the other factors, motivation is a composite of a wide range of sub-factors, however, it is not unreasonable to ask someone with respect to a given scenario: "how motivated would you say you would be to carry out such and such an action, on a scale of zero to five?"

Ability

The third relational factor that is associated with the perceiver is ability. For a person with a physical disability that makes grasping difficult or lifting the arm problematic, the option of carrying anything above the head may simply not be available. In this case, all the other factors may be present, including an umbrella with high tacit capacity and an excellent situated potential, a strong awareness of the umbrella on the part of the perceiver and a correspondingly strong motivation to use it. But inability to grasp the handle renders the affordance zero for this particular person at this particular time.

Preference

The last factor related to the perceiver represents the role played by individual preference. All other factors being equal or even roughly equal, it is often the case that individual adoption of affordances depends at least to some extent on established preferences. In the case of the person who wants to stay dry in the rain, if there are two umbrellas available and one is a favorite, that will probably be the one that gets employed. Preference can be based on any one of a dozen sub-factors, ranging from aesthetic considerations to interpersonal influence to previous personal experience. Preference is distinct, however, from ability, and although preference is related to motivation, the two are not equivalent.

The final factors in the proposed vector space are needed in order to adequately account for features of the situation that are relevant but are not directly related to the relationship between the perceiver and the object. They stand instead for the relationship between the affordance and its context.

Contextual Support

The first of these factors is contextual support, where factors in the environment that are not part of the affordance have an influence one way or the other on the perceiver's interaction with the affordance. There are a wide range of possible contextual supports, including aspects of the situation that are physical, cognitive, and environmental, and the precise nature of the contextual supports in a given situation should be outlined during the process of analyzing the affordance as a whole.

In the example of someone who wishes to stay dry in the rain, the contextual factors would include environmental facts such as how hard it is raining, whether it is warm or cold outside, how hard the wind is blowing and in what fashion, and so on.

Agential Support

The other feature that has not been accounted for yet in an explicit form is the role of other agents in the scenario. Contextual support includes all those factors (excluding the affordance itself) that are present in the environment at the time of the perceiver becoming involved with the affordance. Agential support, on the

other hand, includes those features relating to the roles of the other people, animals, insects, and so on who are also potentially part of the situation. Agents are distinct from other factors of the environment in that they have agency, which is to say volition, goals, and actions of their own, which may have some bearing either directly or indirectly on the particular affordance.

For instance, for the person who wishes to stay dry in the rain, it may turn out that there are other people present who also wish to walk outside. One of them might be elderly or frail and lacking an umbrella, in which case our perceiver could be motivated to behave altruistically and turn over the superior affordance of the umbrella to the other perceiver, choosing instead an inferior solution such as a folded newspaper.

Applying the Model

One straightforward means of applying this model to the study of interface tools is to have participants consider a particular affordance, in order to rate each factor on a Likert scale from 0-5. A rating of zero for any factor effectively zeroes out the strength of the entire affordance, suggesting that it might be further worthwhile to obtain a composite number by multiplication of the individual items. The number could then be used to compare the affordance strength of different kinds of tools. Combining this rating with comments for each factor would add a further layer of information that can contribute to the interface designer's decision process.

References

Gibson, **J. J**. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin.

Morris, M. and Dillon, A. (1997). How User Perceptions Influence Software Use. *IEEE Software*, *14*(4), 58-65.

Turvey, M. T. and Shaw, R. S. (1979). The primacy of perceiving: An ecological reformulation of perception for understanding memory. In L.G. Nilsson (Ed.), *Perspectives on memory research*. Hillsdale, NJ: Erlbaum.

Exploring the Wiki as a Mode of Collaborative Scholarship

Christine RUOTOLO

University of Virginia

Over a decade has passed since Ward Cunningham created his Portland Pattern Repository and coined the term "wiki" (from the Hawaiian word for "quick") to describe the application that enabled its pathbreaking web-based collaborative editing capabilities. Since that time, wikis have flourished on the Web. There are now well over 100 freely available wiki software applications, in a wide variety of programming languages. Wikipedia, the best-known wiki instance, features 2.3 million articles written in over 200 languages by almost 90,000 individual contributors.

Despite this popular success, wikis have been slow to catch on in the academic world, and in the humanities in particular -- surprising given their enormous potential to transform the nature of scholarly communication. Certainly this is due in part to the inherently solitary and linear nature of most academic writing; blogs seem a more intuitive fit for this type of writing, as evidenced by their much greater popularity among scholars. Yet even in the hands of an individual researcher working alone, a wiki can be a valuable tool for organizing and archiving complex scholarly material. In a passage that evokes Vannevar Bush and his mythical memex, Leuf and Cunningham invite scholars to imagine abandoning their scattered scraps of paper and Post-It notes and instead «directly asking your scattered notes where references to 'thingamy' are and having the appropriate bits of paper levitate into view, slide out of bookshelves, and be there at your fingertips ... eventually everything starts to interconnect: notes, files, e-mail, contacts, comments, relational cross-links, Internet resources, and so on ... the whole thing evolves almost organically in response to your growing body of notes" (84).

Of course, the real fun with wikis begins when a complex knowledge representation like the one described above is opened up to a large community of contributors. By virtue of its inherent design, the wiki breaks down the process of text creation into component functions

(structuring, drafting, revising, annotating) typically performed by a solitary author, and allows these to be performed by different persons. Wikis are highly unique in this regard. With blogs and discussion boards, earlier postings cannot be altered except by an administrator and the chronological or topical ordering of the content cannot be restructured. With wikis, however, content is continuously reorganized and revised, often by many different hands. This dispersed, depersonalized method of content creation explains certain rhetorical features of the typical wiki article, which is usually impersonal, unsigned, POV-neutral prose written in the 3rd person. When individual comments and suggestions about a wiki article, often signed and highly personal, are debated, voted upon, and gradually incorporated into the article itself, we see a polyphony of opinion slowly coalesce into a kind of collective understanding. The ideal wiki article would be a highly polished nugget of scholarly consensus, carefully crafted and continuously updated by an informed community, embedded in a dense associative web of related content.

Clearly, there are roadblocks on the way to this ideal. A wiki must achieve a certain critical mass of content before it is truly useful, and without a dedicated core of regular contributors at the outset it is apt to languish unused. "Document mode" -- the anonymous, impartial style of the typical wiki article -- may be familiar from encyclopedias and other reference texts, but it is largely alien to writing in the humanities. Because the precise expression of ideas in words is the very essence of their work, humanities scholars may be particularly reluctant to subject their writing to revision by strangers. Conversely, they are more sensitive to the "legitimation crisis" that arises when text is not attributed in an obvious way (Barton). These problems might exist even if the wiki was restricted to a closed and vetted group of scholarly peers.

The final paper will explore these issues through a series of case studies, and will suggest some optimal uses for wiki technology in an academic setting.

References

Barton, Matthew D. *"The Future of Rational-Critical Debate in Online Public Spheres."* Computers and Composition 22:2 (2005), 177-190.

Bush, Vannevar. *"As We May Think."* The Atlantic Monthly, July 1945.

Dennis, Brian, Carl Smith, and Jonathan Smith. *"Using Technology, Making History: A Collaborative E xperiment in Interdisciplinary Teaching and Scholarship."* Rethinking History 8:2 (June 2004), pp. 303-317.

di Iorio, Angelo and Fabio Vitali. *"Writing the Web."* Journal of Digital Information 5:1 (2004). http://jodi.tamu.edu/Articles/v05/i01/DiIorio

Leuf, Bo and Ward Cunningham. The Wiki Way: Quick Collaborations on the Web. Reading, MA: Addison-Wesley Longman, 2001.

Noel, Sylvie and Jean-Marc Robert. *"How the Web Is Used to Support Collaborative Writing"*. Behavior and Information Technology 22:4 (2003), pp. 245-262.

Obendorf, Hartmut. *"The Indirect Authoring Paradigm – Bringing Hypertext into the Web."* Journal of Digital Information 5:1 (2004). http://jodi.ecs.soton.ac.uk/Articles/v05/i01/Obendorf/

Szybalski, Andy. "Why it's not a wiki world (yet). 14 March 2005. http://www.stanford.edu/~andyszy/papers/wiki_world.pdf

Many Houses, Many Leaves: Cross-Sited Media Productions and the Problems of Convergent Narrative Networks

Marc RUPPEL

Textual and Digital Studies,
University of Maryland College

Humanities scholarship has traditionally viewed a literary work of art as an act of production belonging to larger social and cultural networks, yet remaining relatively fixed within a single medium. The recent identification of patterns of convergence in technological, social/organic, economic and global contexts seems to suggest, however, that although these same traditional models of literary production still constitute a significant portion of the cultural output, they are being transformed and shifted in order to accommodate increasingly intersectional exchanges between media forms and content. In many cases, these shifts have made it possible to develop new structures that shatter the fixity of narrative as a single-medium endeavor and establish instead a multiply-mediated storyworld, a cross-sited narrative, defined here as multisensory "clustered" or "packeted" stories told across a divergent media set. The proliferation of cross-sited narratives—across film, literature, music, video games, live performance and the internet—presents significant challenges to the current modalities of humanistic theory and practice. As both a product of and a reaction to the process of the discrete nature of digitization, cross-sited narratives require us to not only "imagine an infinitely segmentable media market" (Coit-Murphy 91) but also, it seems, an infinitely segmentable and infinitely mediated story that, as a network, draws and exchanges narrative information from site to site. Evidence of this sort of networking can be seen in works such as Mark Danielewski's House of Leaves, which operates across no less than 5 media channels (novel, novella, live performance, recorded music, the web), each integral to the establishment of the narrative storyworld.

The question posed by this paper, then, is whether we can use current models of digital archiving and editioning as the means through which to preserve and distribute a narrative network such as Danielewski's. Is it possible, in the context of contemporary textualities, to retain even a semblance of such a work? Although similar crisis points have always plagued the arts, exposing in many ways the utter ephemerality of even venerated and "durable" technologies such as codex book, we've looked towards the digital as the means through which the fragile materials of paper and print are hardened and made permanent from a bitstream that flows from an electronic fountain of youth. Paradoxically, studies of new media have recently (and perhaps belatedly) moved toward the preservation of digital objects such as early computer games and interactive fictions, recognizing rightly that an entire generation of artifacts is in danger of being obliterated by hardware and software advancements. Although it is universally acknowledged that there is obviously no possible way to replicate the historical moment during which a given work is produced, we have, for the most part, been content to instead refashion the work into whatever single medium seemed to have the greatest potential for preservation. Friedrich Kittler's assertion that "the transposition of media is always a manipulation and must leave gaps" (1990:267) does hold some sway here but, in the sense that a given text (such as Emily Dickinson's letters or William Blake's illuminated manuscripts) has content that is somewhat extractable from its form, such gaps are acceptable when the tradeoffs are vastly improved distribution and conservation. It will be argued here that such an approach will inevitably fail when confronted with the preservation of a cross-sited narrative, as the transcoding of a narrative that relies on the tensions between multiple media sites into a single medium will irrevocably disrupt the network that constitutes its storyworld. We've found ways to overcome the displacement of a medium, but can we hope to approximate a narrative/media network? In short, we can't.

Drawing upon (cognitive) narratology (Herman 2004; Ryan 2003, 2004; Bortolussi and Dixon 2004), cognitive linguistics (Turner 2003; Fauconnier and Turner 2002), convergence theory (Jenkins 2004, 2006; Kittler 1990), archiving strategies for "network" fictions (Montfort and Wardrip-Fruin 2004; Liu, Durand, et al. 2005) and Pierre Levy's theories of virtualization (1998), the focus of this paper will be as follows: 1) to outline the

structures of cross-sited narratives, focusing particularly on their network structure, 2) to assess the current methods of archiving/ preserving literature, such as those proposed by the Electronic Literature Organization, and especially methods that deal with transient texts and networked stories (such as hypertext and interactive fiction) 3) to propose that such methods are inadequate for transcribing the complex interactions between media that occurs within cross-siting and 4) to suggest a new model of temporal textuality that argues that these networks cannot be transposed except through primary materials, and that, often, this primacy is fleeting and not reproducible. In fact, it is possible that the only remnant of these textual networks that will remain are in the annotations and collaborations left by users on message boards, blogs and chat rooms.

The structures that will be recommended for cross-siting are presented as a series of gradient models representing the continuum between a text's materiality and it's narrative. The works that will be studied include Danielewski's House of Leaves, The Matrix, Neil Young's Greendale and the alternate reality game, I Love Bees, a selection that provides a range of blindspots in current practices of textual preservation. Through text, gaming, comic books, live phone calls, the texture of paper and the spontaneity of live performance, each of these works exposes the single-medium logic through which most textual preservation operates. If we are truly entering an era of convergence where media come together in conversation over narrative, then we must also be aware that this coming together is not without consequence. Indeed, the end product of this convergence just might be the erasure of many of the networks it produced.

References

1. **Bortolussi, Marisa and Peter Dixon**. Psychonarratology. Cambridge: Cambridge UP, 2003.
2. **Fauconnier, Gilles and Turner, Mark**. The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. New York: Basic, 2002.
2. **Herman, David.** *"Toward a Transmedial Narratology,"* in Ryan, Marie-Laure (ed.) Narrative Across Media. Cambridge: MIT Press, 2004 (47-75).
3. **Kittler, Friedrich.** Discourse Networks 1800/1900. Stanford: Stanford UP, 1990.
4. **Levy, Pierre.** Becoming Virtual: Reality in the Digital Age. New York: Plennum, 1998.
5. **Liu, Alan , David Durand, Nick Montfort, Merrilee Proffitt, Liam R. E. Quin, Jean-Hugues Réty, and Noah Wardrip-Fruin.** *"Born-Again Bits: A Framework for Migrating Electronic Literature."*
6. **Montfort, Nick and Noah Wardrip-Fruin.** *"Acid-Free Bits: Recommendations for Long-Lasting Electronic Literature."*
7. **Murphy, Priscilla Coit.** *"Books Are Dead, Long Live Books."* inThorburn, David and Henry Jenkins, eds. Rethinking Media Change. Cambridge: MIT Press, 2004 (81-93).
8. **Ryan, Marie-Laure.** Narrative As Virtual Reality: Immersion and Interactivity in Literature and Electronic Media (Parallax: Re-Visions of Culture and Society). Baltimore: Johns Hopkins UP, 2003.
9. ---. Narrative Across Media: The Languages of Storytelling. Lincoln: University of Nebraska Press, 2004.
10. **Turner, Mark.** *"Double-Scope Stories"* in David Herman ed. Narrative Theory and the Cognitive Sciences. Stanford: CSLI, 2003 (117-142).

Can I write like John le Carré?

Jan RYBICKI
Paweł STOKŁOSA

Institute of Modern Languages, Pedagogical University, Krakow, Poland

In a presentation at a previous ALLC/ACH conference, later developed in a paper for *Literary and Linguistic Computing* (Rybicki 2005), one of the authors has begun to investigate the relationship between a literary original and its various translations, basing on Burrows's well-established method first used in his study of Jane Austen (1987) and later developed, evaluated and applied by a number of scholars, including Hoover (2002). Although the results obtained with Delta (Burrows 2002) seem equally promising for computer-assisted translation studies, the first author of this paper (himself a translator of British, American, and Canadian literature) feels that the potential of the older method has not been exhausted in this particular domain – and that it is especially well-suited for case studies such as the one presented here.

The above-mentioned first study investigated character idiolects in a classic Polish trilogy of 19th - century historical romances and its two English translations (made a century apart) and found that many relationships ("distances") between characters in the original were preserved in both, or at least one, of the translations. This time, the works chosen were two "most literary spy novels" by John le Carré, *A Perfect Spy* (1986) and *Absolute Friends* (2003). Written 17 years apart, they were translated by Rybicki into Polish less ten months one from the other in 2003 and 2004.

From the very start, it was evident for the translator that the two novels will be an interesting subject of study due to their being built according to a very similar model, especially where characterization is concerned. Both feature a slightly foolish British agent (le Carré's famous trademark), his highly intellectual yet physically handicapped East German nemesis, the British agent's boss/friend, etc. Since these two very similar works shared their Polish translator – who continued to experience a very strong feeling of *déjà vu* while working on the two novels, this case seemed perfect for a study of stylistic relationships between original and translation.

The following observations have been made: (1) In the narrative, the styles of the originals are more unique than those of the translation. This may be a consequence of the 17 years' distance between the English versions as opposed to the 10 months that separate the translations. (2) Of the three above-mentioned couples of corresponding characters, two are very expectedly similar, while one (the two East-German double agents) are not. Their similarity is "regained" in the translation – an interesting corroboration of the translator's "intuitive" suspicion during his work on the Polish version.

These results show that, at least in this – very special – case, the accuracy of studies performed by Multidimensional Scaling of correlation matrices of relative frequencies of the most frequent words is quite considerable when applied to translation. This is true despite the disquieting fact, to quote Hoover's statement given in a somewhat different context, that "like previous statistical authorship attribution techniques, (this correspondence) lacks any compelling theoretical justification" (2004). The tentative explanations proposed so far by Opas and Kujamaki (using the van-Leuven-Zwart postulate of microstructural changes influencing the text's macrostructure, 1995) or McKenna, Burrows and Antonia ("common words influence syntactic structures and translations of them can influence the meanings we read in a text", 1999), are certainly not enough. Even more telling is the silence that Burrows maintains on the subject in his most translation-oriented study (2002), already quoted above. Since overlapping semantic fields of the most frequent words of texts and divergent linguistic systems make one-on-one correspondences impossible, a more general underlying mechanism must be found. At the same time, empirical studies that have hinted at the existence of such a mechanism have still been very few. This is why more are needed to explain the compelling yet somewhat mystical successes of Burrows's "old" method, and Delta and its various clones. The results presented in this paper are at least a good incentive to study this phenomenon in translations by the same translator; they are, in fact, part of a greater project to investigate the stylometry of all Polish versions by Rybicki, covering a wide range of modern English-language literature.

References

Burrows, J.F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Burrows, J.F. (2002). *The Englishing of Juvenal: computational stylistics and translated texts*. Style 36.

Hoover, D.L. (2002). *New Directions in Statistical Stylistics and Authorship Attribution. Proc. ALLC/ACH*.

Hoover, D.L. (2004) *Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method. Proc. ALLC/ACH*.

McKenna, W., Burrows, J.F. and Antonia, A. (1999). *Beckett's Trilogy: Computational Stylistics and the Nature of Translation, Revue informatique et statistique dans les sciences humaines 35*.

Opas, L.L. and Kujamaki, P. (1995). *A Cross-linguistic Study of Stream-of-consciousness Techniques, LLC* 10/4.

Rybicki, J. (2006). *Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations. LLC*.

What is text? A Pluralistic Approach.

Patrick SAHLE

Universitätsbibliothek Göttingen

Background

Since the beginning of Humanities Computing, text theory has been a fundamental issue. The question „What is text?" has been repeatedly raised at ALLC/ACH conferences. One of the answers, "text is an ordered hierarchy of content objects" (the OHCO-model), has been dominating the discussion for a while and has laid ground for well-established and paradigmatic standards such as the TEI Guidelines. But there has also been criticism since the beginning which questioned this model from different perspectives (Buzzetti, McGann, Olsen, Caton, Huitfeld). Do we in fact have an unsolvable problem? As Jerome McGann puts it: "What is text? I am not so naïve as to imagine that question could ever be finally settled. Asking such a question is like asking 'How long is the coast of England?'."

Starting Points.

Questions are there to be answered. I'll start with the authorities. First, I'll introduce Isidore of Seville, the patron saint of the Internet, who described "oratio" as threefold. Second, I'll take up Willard van Orman Quine's famous ontological slogan "no entity ithout identity" and use it as a lever. Using as a guiding question "When are two things one text?", I will develop a pluralistic theory of text which will integrate a wide range of traditional notions of text. This theory will give us the means to describe the identity conditions of text.

The Wheel of Text.

In a rather free reading, Isidor distinguishes the three components "sense", "expression" and "sign". I relate Isidor's components to the notions of text as meaning and intention (text-i), text as speech and

linguistic utterance (text-l), and text as object and document (text-d). There are other notions of text, which fill the space between those concepts. We talk about the text as structure or 'work' (text-w), placed between text-i and text-l. We also talk about the text as fixed written version (text-v) between text-l and text-d. And we talk about the semiotic text as a picture and as a complex sign (text-s) between text-d and text-i. This concludes the wheel of text and constitutes a comprehensive theoretical model of "What text really is".

Practical implications.

But what is such a theory good for? It has at least two fields of application. First, it has applications in the field of history and in analytical fields. The pluralistic model gives us a tool to describe and locate historical text technologies in relation to certain notions of text. All text technologies promote certain concepts of text while hindering others. This can be shown for oral tradition, manuscript culture and print; for electronic technologies like "plain text", "WYSIWYG" or "markup"; for concepts like hypertext or the OHCO-model; for academic trends like the linguistic turn, pictorial turn or material philology; and even for the description of certain current cultural developments.

In a second application, the pluralistic theory of text has consequences for the assessment and development of future text-technologies. How can markup languages be understood in the light of a pluralistic text model? How are epistemological and ontological questions in markup theory (like the foundational distinction between text and markup) to be answered? How can markup be used and maybe even conceptually extended to truly cover and support *all* notions of text? How can we develop text technologies *beyond* the markup paradigm to cope with notions of text constantly hindered by the concept of markup and the OHCO-model?

Conclusion.

How long is the coast of England? The answer lies in a closer examination of the question: "What do you mean by 'coast'?", "How close will you look at it?" and "Which instruments will you measure it with?". The same holds true for the text. The answer lies in the eye and in the mind of the reader. What is text? - "Text is what you look at. And *how* you look at it."

References

Buzzetti, Dino: *Digital Representation and the Text Model*. In: New Literary History 33 (2002), p. 61-88.

Caton, Paul: *Markup's Current Imbalance*. In: Markup Languages – Theory and Practice 3/1 (2001), p. 1-13.

Ciotti, Fabio: *Text encoding as a theoretical language for text analysis*. In: New Media and the Humanities: *Research and Applications*. Ed. by Domenico Fiormonte und Jonathan Usher. Oxford 2001. p. 39-48.

Dahlström, Mats: *När är en text?*. In: Tidskrift för Dokumentation (Nordic Journal of Documentation) 54/2 (1999), p. 55-64.

DeRose, Steven J.; Durand, David D.; Mylonas, Elli; Renear, Allen H.: *What is Text, Really?* In: *Journal of Computing in Higher Education* 1/2 (1990), p. 3-26.

Durand, David G.; Mylonas, Elli; DeRose, Steven: *What Should Markup Really Be?* Applying theories of text to the design of markup systems. ALLC-ACH'96 Conference Abstracts, University of Bergen, 1996, p. 67-70.

DuRietz, Rolf E.: *The Definition of 'text'*. In: TEXT – Swedish Journal of Bibliography 5/2 (1998), p. 51-69.

Eggert, Paul: *Text-encoding, Theories of the Text, and the 'Work-Site'*. In: Literary and Linguistic Computing 20/4 (2005), p. 425-435.

Flanders, Julia; Bauman, Syd: *Markup, Idealism, and the Physical Text*. ALLC-ACH'04 Conference Abstracts, Göteborg 2004, p. 56-57.

Greetham, David: *The Philosophical Discourse of [Textuality]*. In: Reimagining Textuality: Textual Studies in the Late Age of Print. Ed. by Elizabeth Bergmann Loizeaux und Neil Fraistat. Madison (WI) 2002, p. 31-47.

Hawkins, Kevin; Renear, Allen: Theoretical Issues in Text Encoding: A Critical Review. ALLC/ACH'04 Conference Abstracts, Göteborg 2004, p. 173-175.

Hayles, N. Katherine: *Translating media – Why we should rethink textuality.* In: Yale Journal of Criticism 16/2 (2003), p. 263-290.

Hockey, Susan; Renear, Allen; McGann, Jerome:

What is Text? A Debate on the Philosophical and Epistemological Nature of Text in the Light of Humanities Computing Research. ACH/ALLC'99 Charlottesville (VA).

Huitfeld, Claus: *Multi-Dimensional Texts in a One-Dimensional Medium*. In: Computers and the Humanities 28 4/5 (1995), p. 235-241.

McGann, Jerome: *Endnote: what is text?* In: Ma(r)king the Text – The presentation of meaning on the literary page", ed. by Joe Bray, Miriam Handley und Anne C. Henry. Aldershot u.a. 2000, p. 329-333.

Lavagnino, John: *Completeness and Adequacy in Text Encoding*. In: The Literary Text in the Digital Age. Ed. by Richard Finneran. Ann Arbor (Mi) 1996. p. 63-76.

Phelps, C. Deirdre: *Where's the Book?* The Text in the Development of Literary Sociology. In: TEXT – An Interdisciplinary Annual of Textual Studies 9 (1996), p. 63-92.

Piez, Wendell: *Beyond the "descriptive vs. procedural" distinction*. In: Markup Languages – Theory and Practice 3/2 (2001), p. 141-172.

Renear, Allen: *The Descriptive/Procedural Distinction is Flawed*. In: Markup Languages: Theory and Practice 2/4 (2000), p. 411-420.

Renear, Allen: *Out of Praxis: Three (Meta)Theories of Textuality.* In: Electronic Text - Investigations in Method and Theory. Ed. by Kathryn Sutherland. Oxford 1997. p. 107-126.

Ricœur, Paul: *What is a text? Explanation and understanding*. In: Paul Ricœur, Hermeneutics and the human sciences, Essays on language, action and interpretation. Ed. by John B. Thompson. Cambridge 1981, p. 145-164.

Schreibman, Susan: *Computer-mediated Texts and Textuality: Theory and Practice*. In: Computers and the Humanities 36/3 (2002), p. 283-293.

Smiraglia, Richard P.: *The Nature of "A Work"*. Lanham (Maryland) 2001.

Sutherland, Kathryn: *Revised Relations? Material Text, Immaterial Text, and the Electronic Environment*. In: TEXT – An Interdisciplinary Annual of Textual Studies 11 (1998), p. 35-36.

El Reconocimiento Automático de la Composición en Español.

Octavio SANTANA SUÁREZ
Dept. Informática y Sistemas. Universidad de Las Palmas de G

Francisco Javier CARRERAS RIUDAVETS
Dept. Informática y Sistemas. Universidad de Las Palmas de Gran Canaria

José Rafael PÉREZ AGUIAR
Dept. Informática y Sistemas. Universidad de Las Palmas de Gran Canaria

Virginia GUTIÉRREZ RODRÍGUEZ
Dpt. Estadística, Investigación Operativa y Computación. Universidad de La Laguna

It deals with computerizing one of the processes of words formation in Spanish: the composition. They will solely be studied those cases in which the compound word has been consolidated like the graphical union of the elements that compose it, in regular or irregular way. The formation rules and the application criteria in each case are deduced, consequently, they allow the automated identification of the compound words. The different compounds are extracted from several lexical sources and the applied mechanisms of recognition will be studied, likewise the grammatical categories of original words and the resultant compound. The found recognition criteria are classified and the detected exceptions and irregularities are considered.

INTRODUCCIÓN

La creatividad léxica, según Merving Lang (Lang, 1997), representa una característica fundamental para el habla y la escritura. Los escritores siempre han ideado sus palabras para librarse de las restricciones que les vienen impuestas por el léxico establecido, por lo que utilizan la derivación y la composición como recursos léxicos. Los ejemplos referentes a la formación de palabras se encuentran también en los neologismos de la

terminología científica, en la tecnológica, en el comercio, en los medios de comunicación, en el lenguaje creativo de la literatura moderna y en el lenguaje coloquial e innovador del habla actual. Este trabajo centra su estudio en la yuxtaposición y se excluyen otros por no haberse consolidado como palabra el compuesto resultante —se destaca la importancia de la composición constituida por un elemento verbal y un complemento por ser el más caudaloso de los tipos de composición. Se trata, en suma, de procedimientos para crear neologismos —constituyen una alternativa moderna que enriquece la lengua.

LA COMPOSICIÓN EN ESPAÑOL

La Real Academia de la Lengua Española define la composición como el "procedimiento por el cual se forman palabras juntando dos vocablos con variación morfológica o sin ella —cejijunto, lavavajillas. Se aplica también a las voces formadas con vocablos de otras lenguas, especialmente del latín y el griego —neuralgia, videoconferencia" (RAE, 2001). La composición se sirve de procedimientos para la creación de nuevas palabras, como son: sinapsia, disyunción, contraposición, yuxtaposición, elementos compositivos y acortamiento.

La unión de los miembros en la sinapsia es de naturaleza sintáctica, no morfológica, por lo que es difícil determinar si se ha producido lexicalización o no; suele existir un nexo de unión entre las dos palabras que dan lugar al nuevo término, generalmente con las preposiciones 'de' y 'a' —pan de azúcar, paso a nivel, cuerda sin fin, flor de la abeja. La disyunción da origen a un tipo de lexías en la que los dos elementos participantes no se han soldado gráficamente, por más que la lexicalización sea un hecho —alta mar, peso pluma, pájaro mosca; algunas de tales composiciones pueden llegar a la unión gráfica de sus elementos: caballo de mar-->caballo marino, tela de araña-->telaraña, agua nieve-->aguanieve,... En un grado más alto de unión gráfica está la contraposición, donde los elementos que participan se escriben unidos por un guión que, generalmente, no aparecerá debido a las restricciones del español —coche-bomba-->coche bomba, faldapantalón-->falda pantalón—, aunque la Real Academia Española establece que "cuando no hay fusión sino oposición o contraste entre los elementos componentes, se unirán estos con guión" (RAE, 1995) —físico-químico. El más generoso de los procesos de composición es la yuxtaposición o lexías compuestas, aquí la fusión gráfica de los elementos participantes en el compuesto es total, así como su lexicalización y su gramaticalización —carnicol, malqueda, cochitril, hincapié. Aunque la frontera entre derivación y composición no resulta muy clara, sobre todo en el caso del abreviamiento —coyotomate— o la acronimia —información automática-->informática, poliestar galo-->tergal—, muchos autores consideran el acortamiento como un procedimiento de formación de neologismos que por su naturaleza no constituiría una derivación sino que más bien formaría parte de la composición. La utilización de raíces cultas greco-latinas es frecuente en los procesos de generación de nuevas palabras —particularmente en los campos científicos y técnicos—; las voces en cuya formación intervienen podrían, según varios autores, no considerarse propiamente compuestas, ya que la mayoría de sus raíces no pueden aparecer aisladamente, pero tampoco pueden considerarse derivadas, puesto que tienen un comportamiento peculiar —significado léxico— que los aleja de los auténticos afijos. A este tipo de raíces se les da el nombre de elementos prefijales o sufijales —elementos compositivos—, en función de si se anteponen a otra raíz o se posponen.

En el presente estudio se tratan, desde un punto de vista morfológico, los compuestos yuxtapuestos o lexías compuestas, al igual que algunos casos especiales de acortamiento, elementos compositivos y parasíntesis por composición. Los restantes tipos no se consideran debido a la dificultad para justificar que constituyen un verdadero compuesto en español, ya que habría que tener en cuenta factores sintácticos y semánticos que inicialmente no se consideran en este trabajo.

REGLAS DE COMPOSICIÓN

Se parte de una base de unos 4000 compuestos recopilados del Diccionario General de la Lengua Española Vox (Bibliograf, 2003) y del glosario de compuestos del libro "La composición nominal en español" de Eugenio Bustos (Bustos, 1986) —basado en obras de carácter general, DRAE, y en otras de carácter regional o dialectal: hablas leonesas, aragonesas, meridionales, español de América—, además, se han añadido unos 6000 compuestos, que incorporan elementos prefijales, procedentes de diversos diccionarios de español (Clave, 1997; Espasa Calpe, 1991; Casares, 1990; Larousse, 1996; Alvar, 2003; Moliner, 1996;). Los compuestos analizados se clasifican en grupos según la categoría gramatical de sus constituyentes.

Se busca, a partir del estudio del comportamiento de los vocablos constituyentes del compuesto, las reglas de formación del mismo; algunas coinciden con las tratadas por algunos lingüistas, aunque con una adaptación informática justificada por el comportamiento mayoritario observado —aeriforme-->aeri + forme, según el Diccionario General de la Lengua Española Vox (Bibliograf, 2003), sin embargo, el comportamientos mayoritario es aero. Se define regla de formación a todo comportamiento mayoritario que permita concretar un mecanismo capaz de relacionar los elementos constituyentes del compuesto, para su reconocimiento por medios informáticos. Se estudian además, las reglas fonéticas —cambios gráficos para mantener el sonido de una consonante: anquirredondo-->anca + redondo— que se producen como consecuencia de haber aplicado una regla de formación. Se obtiene un conjunto de reglas que, junto a las excepciones encontradas, permiten el reconocimiento automático de las palabras compuestas y en el futuro su generación.

PROCESAMIENTO DE LAS REGLAS DE COMPOSICIÓN

Se parte de la palabra compuesta y se comprueba que cumpla unas ciertas condiciones —tamaño de la palabra, mayor a cinco caracteres, o bien, número de sílabas, mayor que tres: uñalbo. El proceso de reconocimiento propone partir la palabra hasta que se encuentre un vocablo o ambos, a los que se les aplica la regla correspondiente o bien se tratan como excepción; se pueden obtener múltiples soluciones —algunas o todas incorrectas. La secuencia de cortes permite añadir otro tipo de condicionantes: por ejemplo, las palabras que forman el compuesto no deben ser derivadas, sino constituir una unidad léxica —no contener prefijos, particularmente en el primer elemento del compuesto—, o no admitir la flexión del diminutivo en la segunda palabra del compuesto.

Hay que tener en cuenta que en un estudio cuyo objetivo sea la automatización de la composición con medios informáticos, los aspectos formales o teóricos no tienen por qué coincidir con los estrictamente lingüísticos. Así, *clarovidente —falsa composición, pues lo correcto sería clarividente— no tendría por qué tratarse de una mala formación al no contravenir ninguna regla fonotáctica del lenguaje, ni siquiera la norma de la estructura silábica del español.

CONCLUSIONES

Se trata de un trabajo novedoso, ya que han resultado infructuosas las búsquedas de referencias sobre procesamiento automático de la composición en español, a pesar de la presumible trascendencia de tal proceso.

Internet y el lenguaje periodístico recogen, con frecuencia, neologismos compositivos debido a la rápida evolución de los acontecimientos y a su inmediata trascripción al mundo de las tecnologías de la información: movichandal, ciberamor, eurosueldo,... Son imprescindibles procesos automáticos que sean capaces de identificar estas palabras y situarlas en un contexto lingüístico adecuado: morfológico y semántico.

El reconocimiento de palabras compuestas en español es útil en aplicaciones para el procesamiento automático del lenguaje natural, debido a que lleva implícito vínculos semánticos, sobre todo en los compuestos endocéntricos. Asimismo, potencia las búsquedas en Internet al ampliar el abanico de relaciones morfoléxicas deducidas de los compuestos estudiados, sus derivaciones y flexiones

Referencias

1. **Biblograf, s.a.** 2003. *"Diccionario General de la Lengua Española VOX"* en CDROM. Barcelona.
2. **Clave SM.** 1997. *"Diccionario de Uso del Español Actual"*. Clave SM, edición en CD ROM. Madrid.
3. **David Serrano Dolader.** 1995. *"Las formaciones parasintéticas en español",* Ed. Arco/Libros, S.L.
4. **Espasa Calpe.** 1991. *"Gran Diccionario de Sinónimos y Antónimos",* 4a edic. Espasa Calpe, Madrid.
5. **Eugenio Bustos Gisbert.** 1986. *"La composición nominal en español",* Universidad de Salamanca.
6. **Jose Alberto Miranda**. 1994. *"La formación de palabras en español"*. Ediciones Colegio de España.
7. **Julio Casares. 1990.** *"Diccionario Ideológico de la Lengua Española",* 2a Edición. Ed. Gustavo Gili, s.a. Barcelona.
8. **Larousse Planeta, s.a.** 1996. *"Gran Diccionario de la Lengua Española"*. Larousse Planeta, s.a., Barcelona.

9. **Manuel Alvar Ezquerra.** 2002. *"La formación de las palabras en español"*. Cuadernos de lengua española, Ed. Arco/Libros, Madrid.

10. **Manuel Alvar Ezquerra.** 2003. *"Nuevo diccionario de voces de uso actual"*. Ed. Arco/Libros, Madrid.

11. **María Moliner.** 1996. *"Diccionario de Uso del Español"*, edición en CD ROM. Gredos, Madrid.

12. **Mervyn Francis Lang.** 1992. *"Formación de palabras en español. Morfología derivativa productiva en léxico moderno"*. Cátedra, Madrid.

13. **Ramón Almela Pérez.** 1999. *"Procedimientos de formación de palabras en español"*. Ed. Ariel Practicum.

14. **Real Academia Española y EspasaCalpe**. 2001. *"Diccionario de la Lengua Española"*, edición electrónica. 22a edn. Madrid.

15. **Soledad Varela Ortega.** 1990. *"Fundamentos de Morfología"*, Ed. Síntesis.

16. **Waldo Pérez Cino.** 2002. *"Manual Práctico de formación de palabras en español I"*, ed. Verbum.

A Fresh Computational Approach to Textual Variation

Desmond SCHMIDT
School of Information Technology and Electrical Engineering
University of Queensland

Domenico FIORMONTE
Università Roma Tre

If there is one thing that can be said about the entire literary output of the world since the invention of writing it is that literary works exist in multiple versions. Such variation may be expressed either through the existence of several copies of a work, through alterations and errors usually in a single text, or by a combination of the two. A textual feature of this degree of importance ought to be at the forefront of efforts to digitise our written cultural heritage, especially at a time when printed media are becoming less important. Until now literature has been represented digitally through systems of markup such as XML, which are ultimately derived from formal languages developed by linguists in the 1950s (Chomsky 1957; Hopcroft and Ullman 1969); but over recent years it has gradually become clear that the hierarchical structure of such languages is unable to accurately represent variation in literary text. Alan Renear (1997), for example, admits that variation is one exception that does not fit into his hierarchical model of text; likewise Vetter and McDonald (2003) conclude that markup provides 'no entirely satisfactory method' for representing variation in the poetry of Emily Dickinson. More general discussions of the shortcomings of hierarchical markup, including the problem of variation, have recently been made by Dino Buzzetti (2002) and Edward Vanhoutte (2004).

An alternative approach, not yet tried, is to use graphs to represent variation. Graphs were first studied in the 18th century by the Swiss mathematician Leonhard Euler, who is best remembered for his solution to the famous 'Bridges of Königsberg' problem (Trudeau 1993). The type of graph which most closely resembles textual variation does not appear to have yet been described by

anyone; however, it can be derived from the following example. Consider four versions of the simple sentence:

A The quick brown fox jumps over the lazy dog.

B The quick white rabbit jumps over the lazy dog.

C The quick brown ferret leaps over the lazy dog.

D The white quick rabbit playfully jumps over the dog.

Collapsing the five versions into collapsing the four versions. Such repetitions are clearly undesirable. If they were present in an electronic edition each time one copy was changed, an editor would have to check that the other copies were changed in exactly the same way. If all this redundancy is removed by collapsing the four versions wherever the text is the same, the following graph results:

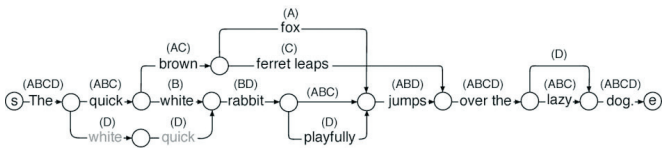


Figure 1

This is a type of 'directed graph', which we call a 'textgraph'. Its key characteristics are:

a. It has one start and one end point.

b. The 'edges' or 'arcs' are labelled with a set of versions and with a fragment of text, which may be empty.

c. There are no 'directed cycles' or loops.

d. It is possible to follow a path from start to end for each version stored in the graph, which represents the text of that version.

In figure 1 version D contains an insertion: 'playfully' and a deletion 'lazy'. These are represented in the graph as empty edges. In fact insertions and deletions are the same thing viewed from different perspectives: every deletion is an insertion in reverse and vice versa. Transpositions, as in version 'D' - the transposition of 'white' and 'quick' in relation to version 'B', can be viewed as a deletion of some text in one place and its insertion in another. All that is then needed is some way to refer back to the original text to avoid copying, e.g. by 'pointing' to it. This feature has been shown in figure 1 by drawing the transposed text in grey, which does not change the structure of the graph.

This model is equally applicable to variation arising from a single manuscript or from the amalgamation of multiple manuscripts of the same work. Its biggest advantage is that it can handle any amount of overlap without duplicating text. One example of a rigorous test of this model can be found in the archives of the 'Digital Variants' website. The poem 'Campagna Romana' by the modern Italian poet Valerio Magrelli exists in four drafts, the first of which is shown in figure 2.

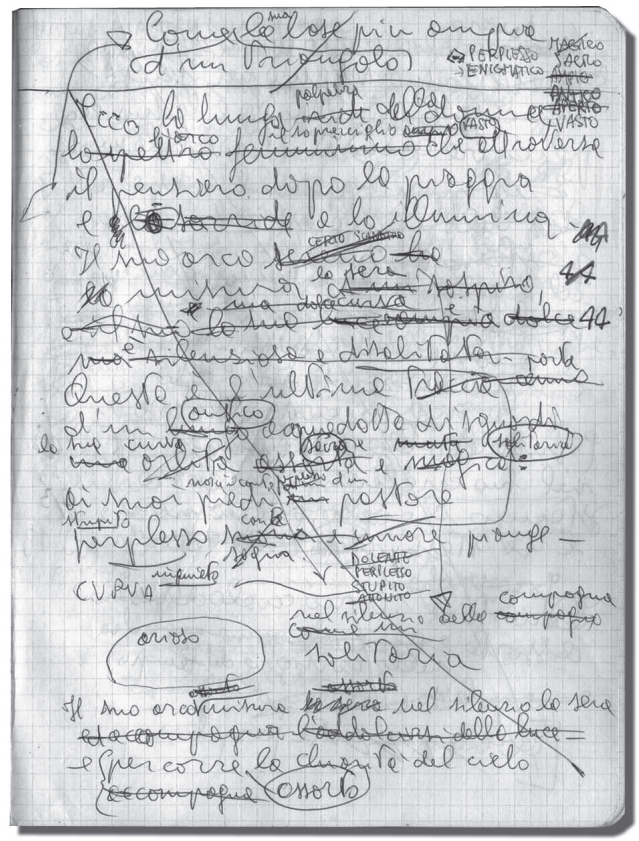


Figure 2

In original manuscripts like this it is often unclear how variants are to be combined. For example, in the line 'Il suo arco sereno/certo/scandito/ ha la misura d'un sospiro/misura la sera/' it is impossible to say if there ever was a version: 'Il suo arco scandito misura la sera'. The sensible way to proceed here is simply to provide

a mechanism for recording any possible set of readings, and to leave the interpretation up to the editor.

Figure 3

Documents which are based on this graph structure we call 'multi-version documents' or mvd's. One application of this format is the applet viewer shown in figure 3 (Schmidt, 2005). This currently allows the user to view one readable version or layer of text at a time. In reality only the differences between each layer are recorded, and the user can highlight these using red to indicate imminent deletions and blue for recent insertions. The text is also searchable through one version or all versions simultaneously. This visualisation tool is in an early stage of development and as yet it can only handle plain text. However, because it cleanly separates the content of the document (represented by the edges of the graph) from its variation (represented by the graph's structure), the same method could also be used to record versioning information in almost any kind of document - including XML, graphical, mathematical and other formats:

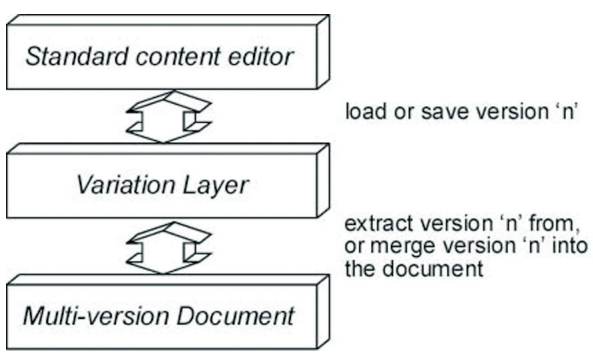


Figure 4

This allows a multi-version document to utilise existing technology. By removing variability from a text, and effectively representing it as a separate layer, the mvd format allows technologies like XML to be used for what they were designed to do: to represent non-overlapping content. One way this could be achieved would be to edit the text in an existing editor but to modify the editor slightly so that instead of reading and writing the document directly it would read and write only one version at a time to an mvd file, as shown in figure 4.

There are a couple of possible objections to the overall technique described here. Firstly, because it is not based on markup, it is no longer practical for the user to see the contents of the document in its merged form. Secondly, existing XML technology currently utilises markup to record information about the status of individual variants. This data would have to be re-encoded as characteristics of the bits of varying text, since the document content would no longer carry any information about variation. However, the very idea of 'variants' embedded in the text is a structure inherited from the critical edition, which is now widely regarded as obsolescent (Ross 1996; Schreibman 2002). Through the printed medium traditional philology advanced the notion of textual 'truth' in its effort to restore a lost original, whereas our model is directed toward the fruition of the text as it really is. As we move forward into an age when digital text has the primary focus, some of the old ideas associated with paper-based methodologies may have to be revised or given up entirely (Fiormonte 2003).

In conclusion, the use of 'textgraphs' to represent variation appears to overcome the problems of redundancy and overlap inherent in current technologies, and to reduce document complexity. Thus far, a file format has been devised and has been demonstrated in a working multi-version document viewer for plain text, which is capable of representing original documents of high variability. By separating variation from content it also has the potential to leverage existing document handling technologies. This technique represents a new method of handling textual variation; it is mathematical and wholly digital in character, and unlike what it purports to replace, it is not based on the inherited structures of the printed edition.

References

Buzzetti, D. (2002) Digital Representation and the Text Model, *New Literary History,* 33(1): 61-88.

Chomsky, N. (1957) *Syntactic Structures,* Mouton & Co: The Hague.

Fiormonte, D. (2003) *Scrittura e filologia nell'era digitale,* Bollati Boringhieri: Turin.

Hopcroft, J.E. and Ullman, J.D. (1969) *Formal Languages and their Relation to Automata,* Addison-Wesley: Reading, Massachusetts.

Renear, A. (1997) Out of Praxis: Three (Meta) Theories of Textuality in *Electronic Text,* Sutherland K. (ed.), Clarendon Press: Oxford, 107-126.

Ross, C. (1996) The Electronic Text and the Death of the Critical Edition in R. Finneran (ed.), *The Literary Text in the Digital Age,* 225-231.

Schmidt, D. (2005) MVDViewer Demo, available at: http://www.itee.uq.edu.au/~schmidt/cgi-bin/MVDF_sample/mvdviewer.wcgi

Schreibman, S. (2002) The Text Ported, *Literary and Linguistic Computing,* 17: 77-87.

Trudeau, R.J. (1993) *Introduction to Graph Theory,* Dover: New York.

Vanhoutte, E. (2004) Prose Fiction and Modern Manuscripts Limitations and Possibilities of Text-Encoding for Electronic Editions in Unsworth, J., O'Brien, K. O'Keeffe and Burnard, L. (eds.), *Electronic Textual Editing*. (forthcoming - available at http://www.kantl.be/ctb/vanhoutte/pub.htm#arttg)

Vetter, L. and McDonald, J. (2003) Witnessing Dickinson's Witnesses, *Literary and Linguistic Computing,* 18: 151-165.

Cross-Collection Searching: A Pandora's Box or the Holy Grail?

Susan SCHREIBMAN
Gretchen GUEGUEN
Digital Collections & Research, UM Libraries

Jennifer O'BRIEN ROPER
Original Cataloging, UM Libraries

While many digital library initiatives and digital humanities centers still create collection-based projects, they are increasingly looking for ways of federating these collections, enhancing the possibilities of discovery across media and different themed-research. Facilitating access to these objects that are frequently derived from different media and formats, while belonging to different genres, and which have traditionally been described in very different ways, poses challenges that more coherently-themed collections may not.

In the last few years it has become increasingly evident to those in the digital humanities and the digital library communities, and the agencies which fund their research, that providing federated searching for the immensely rich digital resources that have been created over the past decade is a high priority. Several recent research grants speak to this issue, such as the Mellon-funded NINES: A Network Funded Initiative for Nineteenth Century Electronic Scholarship , or The Sheet Music Consortium .

While digital objects organized around a specific theme or genre typically provide opportunities for rich metadata creation, providing access to diverse collections that seem to have little in common (except that they are owned by the same institution) often poses problems in the compatibility of controlled vocabulary and metadata schema. While this problem has been noticed on much larger scales before and addressed by initiatives such as z39.50 and the Open Archives Initiative's Protocol for Metadata Harvesting, addressing the problem within a library's or center's own digital collections is a vital part of making such initiatives successful by leveraging cross-collection

discovery through the internal structure of the metadata scheme as well as a consistent approach to terminology. This presentation will explore the issues surrounding creating an archive of cross-searchable materials across a large spectrum of media, format, and genre at the University of Maryland Libraries . It will examine the way some of these interoperability problems can be addressed through metadata schema, targeted searching, and controlled vocabulary.

Description of Research

This paper will be based on the research done at the University of Maryland Libraries using two ongoing projects. The first project utilizes The Thomas MacGreevy Archive , a full-text digital repository (following the Text Encoding Initiative (TEI) Guidelines), to explore the development of metadata and descriptors to facilitate searching across individual collections which are described at different levels of granularity. The second project involves using the knowledge based on the research carried out for the more cohesive MacGreevy Archive for the more diverse repository the UM Library is developing utilizing Fedora as its underlying repository architecture.

The Thomas MacGreevy Archive is being explored as a microcosm from which to examine issues of searchability of content divided into collections that cannot be described using a single controlled vocabulary and has different modes of display. The necessity for cross-collection searches has arisen due to the Archive expanding its content from digitized versions of books and articles, to two collections of correspondence (one relatively small collection of seven letters, the other quite large, c 150 letters), and making images that are currently available only via hyperlink from within texts individually discoverable. Preliminary findings involving this research were shared at the joint 2005 ACH/ALLC Conference at University of Victoria.

Another issue that the MacGreevy Archive can model is problems of controlled vocabularies across different collections. The current controlled vocabulary descriptors use a faceted approach to describe articles and books written on such topics as art, music, and literature. Since both the correspondence and images differ in form and function from the existing objects in the archive, the current controlled vocabulary descriptors are not granular enough to capture either the variety of themes common to letters, or the additional descriptors to describe what the images are of and about.

The experience gained in exploring the more homogenous MacGreevy Archive is being applied to the much more diverse collections and formats being housed in the Fedora repository in which rich collection-specific controlled vocabulary across multiple formats is being developed at the same time as a vocabulary which provides users the opportunity of discovery across all collections. While specific controlled vocabularies exist that would adequately describe each collection, they are generally too specific for materials outside that collection. On the other hand, Library of Congress Subject Headings (LCSH), while sufficiently broad in scope, are unwieldy in form, taking a post-faceted approach by combining several smaller descriptors into a predefined string. These long strings cause multiple problems in searching (including not being amenable to Boolean searching) but are ubiquitous within university libraries, forming the underlying basis for the vast majority of online catalogues. LCSH descriptors will be necessary, however, for those objects that will occur concurrently in the library catalog.

In exploring cross-collection search capabilities in this larger, more diverse environment, the use of controlled vocabulary for subject access is only one possible source of commonalities. Within the Fedora repository, the University of Maryland Libraries will create a metadata scheme that represents a hybrid of the elements and concepts chiefly found in qualified Dublin Core , and the Visual Resources Association Core < http://www.vraweb.org/vracore3.htm >. This scheme and hybrid approach was first used by the University of Virginia, and the UM Libraries is refining that element set and list of required elements specifically with cross-collection searching in mind. By requiring elements that include information such as the century and geographic focus of individual objects, designers aim to define and render searchable the broader topics that objects from disparate collections of narrow focus may have in common. Designers must also use or define standard vocabularies to be used to populate these broader elements to ensure successful cross-collection discovery.

Previous Research

The integrated design of online information retrieval systems has been studied most prominently by Marcia Bates (2001). However, most research in this field

takes a more atomized approach, focusing solely on one aspect of design: metadata schemes for instance, or GUI design for search screens. Other research has examined the particular needs and searching habits of users, particularly in humanities disciplines, when faced with online search interfaces. Prominent among these has been the work of Deborah Shaw (1995) and the series of reports from the Getty Online Searching Project (Bates, 1994, 1996; Bates et.al. 1993, 1995; Siegfried et.al. 1993). The majority of this research was carried out in the late 1990s and follow-up has been more in specific application of digital library systems than in respect to user-oriented, integrated design, such as Broughton (2001).

Another area of inquiry relevant to this research involves the use of faceted classification systems for web-based discovery. The most recent white paper produced by the NINES project neatly summarizes current research (NINES 2005). KM's 'The Knowledge Management Connection' discusses faceted classification within the context of information-intensive business environments. Denton (2003) discusses how to develop a faceted classification scheme, while Bates (1988) surveys the various approaches to subjection description for web-based discovery.

Conclusion

This paper will build on previous research as mentioned above in the development of a controlled vocabulary, metadata schema, and faceted classification scheme which provides for both rich collection-specific discovery, as well as federated searching across collections.

References

Bates, M.J. (1988) *How to Use Controlled Vocabularies More Effectively in Online Searching*. Online 12 (November) 45-56.

Bates, M. J., Wilde, D.N., & Siegfried, S. (1993). *An analysis of search terminology used by humanities scholars: the Getty online searching project report no. 1*. Library Quarterly, 63. (September). 1-39.

Bates, M. J. (1994). *The Design of Databases and Other Information Resources for Humanities Scholars: The Getty Online Searching Project Report No. 4*. Online and CD-ROM Review, 18. (December). 331-340.

Bates, M. J. (1996). *Document Familiarity in Relation to Relevance, Information Retrieval Theory, and Bradford's Law: The Getty Online Searching Project Report No. 5*. Information Processing & Management 32. 697-707.

Bates, M.J. (2002). *The cascade of interactions in the digital library interface*. Information Processing and Management: An International Journal, 38:3. (May). 381-400.

Bates, M. J., Wilde, D.N., & Siegfried, S. (1995). *Research Practices of Humanities Schoars in an Online Environment: The Getty Online Searching Project Report No. 3*. Library and Information Science Research, 17 (Winter). 5-40.

Broughton, V. (2001). *Faceted Classification as a Basis for Knowledge Organization in a Digital Environment; the Bliss Bibliographic Classification as a Model for Vocabulary Management and the Creation of Multi-Dimensional Knowledge Structures*. The New Review of Hypermedia and Multimedia, 7. 67-102.

KM (Knowledge Management Connection). Faceted Classification of Information.

NINES: *A Federated Model for Integrating Digital Scholarship*. (2005) White Paper (September) < http://www.nines/org/about/9swhitepaper.pdf>

Shaw, D. (1995). *Bibliographic database searching by graduate students in language and literature: search strategies, system interfaces, and relevance judgements*. Library and Information Science Research, 17. 327-45.

Siegfried, S., Bates, M. J. and Wilde, D. N. (1993). *A Profile of End-User Searching Behavior by Humanities Scholars: The Getty Online Searching Project Report No. 2*. Journal of the American Society for Information Science, 44. (June). 273-291.

William, Denton. (2003) *How to Make a Faceted Classification and Put it on the Web*.

Exploring the Generic Structure of Scientific Articles in a Contrastive and Corpus-Based PersPective

Noëlle SERPOLLET
Céline POUDAT

University of Orléans, France

This paper will describe and analyse the generic structure of linguistic articles, using a corpus-based methodology and working within a contrastive (English-French) perspective. The main question that we wish to answer is the following: "Do scientific articles – and more particularly linguistics ones – have a generic structure, and to what extent does this structure vary from one language to another?"

We will answer this linguistic question using techniques from computational and corpus linguistics.

The notion of genre is more and more present, as much in linguistics as in information retrieval or in didactics. Genres and texts are intimately connected, as genres could not be tackled within the restricted framework of the word or the sentence. Indeed, genres can only be perceptible using text corpora both generically homogeneous and representative of the genre studied. The progress of information technology and the possibilities of digitization have made it possible to gather homogeneous and synchronic corpora of written texts to analyse and characterize genres.

Moreover, the development of computational linguistics, of linguistic statistics and more generally of corpus linguistics has led to that of tools and methods to process large corpora which make it possible nowadays to detect linguistic phenomena and regularities that could not have been traced before. In that sense, inductive typological methods and multi-dimensional statistical methods (see Biber, 1988) seem crucial to make the criteria which define the genres appear more clearly.

If literary genres have been largely explored, the study of academic / scientific and professional genres has mainly been undertaken for about thirty years within a more applied trend. English for Specific Purpose, is a rhetorical-functional trend which is interested in macro-textual descriptions and in describing genres from a phrasal or propositional point of view. The description of rhetorical moves (see Swales, 1990) is rather qualitative than quantitative, as the moves can scarcely be automatically identified – although several studies have set out to demonstrate their relative identification by training classifiers on manually annotated corpora (Kando, 1999 and Langer et al., 2004).

Our perspective is however different, as we do not start from a set of predefined moves: our objective is indeed to describe the genre of the article and its structure in a quantitative perspective, starting from three levels of description: the structural, the morphosyntactic and the lexical level.

The study is based on a generically homogeneous corpus composed of French and English journal articles that all belong to the linguistic domain, chosen as this is the field we have the best expertise in. The French corpus is made up of 32 issues of 11 linguistic journals, that amounts to 224 articles; whereas the English one includes 100 articles, that is 16 issues of 4 linguistic journals. Texts have all been issued between 1995 and 2001 to limit the possibilities of diachronic variations.

In order to describe the document structure of scientific articles, we first marked up the document structure and the article constituents according to the Text Encoding Initiative Guidelines (Sperberg-McQueen et al., 2001), to ensure the corpus reusability and comparability with other corpora: the article sections were taken into account (introduction, body, divisions, conclusion), as well as its titles, subtitles, and specific components (examples, citations, appendices, etc.).

This XML markup enabled us to obtain the main characteristics of the article structure and organization in the two languages (number of sections, structure depth, etc.) and to assess their stability and differences, using XSL stylesheets.

Once these characteristics were established, we focused on the article sections: as both French and English linguistics articles are not submitted to an IMRAD structure (Introduction, Materials and methods, Results, Analysis,

Discussion), only introductions and conclusions could be directly observed and compared. Indeed, it would have been irrelevant to analyze "third sections" as many texts are only divided into two main parts.

The linguistic properties of introductions and conclusions were described thanks to two different levels of description: the lexical and the morphosyntactic levels, which did not require the same processing. The lexical characteristics of the sections were first obtained using Alceste and its Hierarchical Descendant Classification.

We then concentrated on the morphosyntactic level, on the one hand because morphosyntactic variables easily lend themselves to voluminous data as they are formal enough to be tagged and calculable and on the other hand because various studies have demonstrated their efficiency in genre processing (Karlgren & Cutting 1994; Kessler et al., 1997; Malrieu & Rastier 2001; Poudat, 2003).

Although several taggers are available, they are generally little adapted to the processing of scientific texts; for instance, the French Inalf Institute trained Brill tagger on 19th century novels and *Le Monde* articles. Most of the English taggers are trained on the Penn TreeBank corpus and use very robust tagsets which interest is descriptively very weak. As many available taggers are trainable (Brill Tagger, TreeTagger, TnT tagger, etc.), we decided to develop our own tagset and to generate a new tagger devoted to the processing of scientific texts. We then used a specific tagset of 136 descriptors (described in Poudat, 2004) to process the French corpus. The tagset is devoted to the characteristics of scientific discourse, and gathers the general descriptive hypothesis put forward in the literature concerning scientific discourse. Among the very specific variables we developed, we can mention symbols, title cues (such as 1.1.), modals, connectives, dates, two categories for the IL personal pronoun, in order to distinguish between the French anaphoric and impersonal IL, etc.

The training task is very costly, as it requires the building of a manually annotated training corpus that has to be large enough to enable the system to generate tagging rules. For this reason, it was only led on the French corpus, using TnT tagger. We then adapted the tags and the outputs of CLAWS (the Constituent Likelihood Automatic Word-tagging System developped at Lancaster University, see Garside, 1987) to get comparable data.

The morphosyntactic characteristics of the French and English introductions and conclusions were then determined, using statistical methods.

After having described our methodology, we will present the results obtained thanks to this same methodology. The last part of our paper will discuss the conclusions that could be drawn from these findings.

References

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Garside, R. (1987). The CLAWS Word-tagging System. In Garside, R., Leech, G. and Sampson, G. (eds), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Kando, N. (1999). Text Structure Analysis as a Tool to Make Retrieved Documents Usable. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, Taipei, Taiwan, Nov. 11-12, pp. 126-135.

Karlgren, J. and Cutting, D. (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *Proceedings of COLING 94*, Kyoto, pp. 1071-75.

Kesler, B., Nunberg, G. and Schütze, H. (1997). Automatic Detection of Genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*. San Francisco CA: Morgan Kaufmann Publishers, pp 32-38.

Langer, H. Lüngen, H. and Bayerl, P. S. (2004). Towards Automatic Annotation of Text Type Structure. Experiments Using an XML-annotated Corpus and Automatic Text Classification Methods. *Proceedings of the LREC-Workshop on XML-based richly annotated corpora*, Lisbon, Portugal, pp. 8-14.

Malrieu, D. and Rastier, F. (2001). Genres et variations morphosyntaxiques. TAL, 42(2) : 547-578.

Poudat, C. (2003). Characterization of French Linguistic

Research Papers with Morphosyntactic Variables. In Fløttum K. and Rastier F. (eds). *Academic discourses - Multidisciplinary Approaches*. Oslo :Novus, pp. 77-96

Poudat, C. (2004). Une annotation de corpus dédiée à la caractérisation du genre de l'article scientifique. *Workshop TCAN - La construction du savoir scientifique dans la langue*, Maison des Sciences Humaines - Alpes, Grenoble, 20 octobre 2004.

Sperberg-McQueen, C.M. and Burnard, L. (eds) (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Charlottesville, Bergen: Text Encoding Initiative Consortium. XML Version.

http://www.tei-c.org/Guidelines2/index.xml.ID=P4 and http://www.tei-c.org/P4X/ (accessed 14 November 2005).

Swales, J. (1990). *Genre Analysis : English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Giving Them a Reason to Read Online: Reading Tools for Humanities Scholars

Ray SIEMENS
English, U Victoria
siemens@uvic.ca

John WILLINSKY
Education, U British Columbia

Analisa BLAKE
Geography, U Victoria

Overview

A good deal of the emerging research literature concerned with online information resources focuses on information retrieval, which is concerned with the use of search engines to locate desired information. Far less attention has been paid to how the found materials are read and how that critical engagement can be enhanced in online reading environments. Building on research reported in relation to Willinsky's work in and around the Public Knowledge Project (http://www.pkp.ubc.ca/), Warwick's "No Such Thing as Humanities Computing?" (ALLC/ACH 2004), and Siemens, et al., "The Humanities Scholar in the Twenty-first Century" (ALLC/ACH 2004), among others pertinent to our community, this paper reports on a study examines the question of whether a set of well-designed reading tools can assist a sample of 15 humanities computing scholars in comprehending, evaluating and utilizing the research literature in their area. In progress at present, the study of our group will conclude in January 2006, with results available in March 2006. Should the paper be accepted for presentation at the conference, at that time the authors will revise their abstract, fully detailing the results of the study, which will include an analysis of the degree to which different types of reading tools (providing background, related materials, etc.) contributed, if at all, to these scholars' reading experience.

Context

The larger study, a subset of which is devoted to computing humanists, investigates how journal websites can be designed to better support the reading of research in online settings for a wider range of readers than has traditionally been the case with research. Given that well over 75 percent of research journals now publish online, with a number of them made free to read, the reading experience and audience for research is changing. This work looks at whether the design and structure of the journal's "information environment" can improve the reading experience of expert and novice readers of this literature. Specifically, it examines whether providing far richer context of related background materials for a given text than is available with print, assists the online reading process. In this way, the study seeks to understand, better, reading for information in online environments.

Growing out of Willinsky's design work in online information environments for schools, policy forums, and academic journals over the last five years, and work within the digital humanities, this study evaluates whether the specific online tools and resources that journals are now able to provide can assist expert and novice readers in making greater sense and use of the research literature. By drawing on related work in reading comprehension in schools, as well as from initial design experiments with journals in online settings, it is posited that information environments that provide links to related resources will enable a wide range of readers to establish a greater context for comprehending and potentially utilizing the research they have come to read. It may also support the critical engagement of more expert readers.

This study focuses on testing a context-rich Reading Tools which can accompany each journal article (and if proven useful can also be used with online conference papers, reports, and theses). This tool provides (a) background on the article and author, (b) links from each research article to directly relevant materials (based on the keywords provided by the author), and (c) opportunities for interactivity, such as commenting and contacting the author. Utilizing research studies in medicine and education as the publishing content to be read online, the contribution of this Tool will be assessed with a sample of faculty members and students in education and medicine, as well as well as with a sample of policymakers and members of the public.

The study, in this case, is being conducted with computing humanities scholars. A sample of 15 scholars will read an article in their field, and utilize the Reading Tools to see which tools, if any, and to what degree, these tools contribute to their comprehension, evaluation, and interest in utilizing the work they are reading. These tools will connect the reader to the author's other works, related studies, book reviews, summaries of literary critics; work, online forums, instructional materials, media reoports, and other data bases. The readers will be asked to reflect on their reading and use of the tools in a think-aloud protocol, with the researcher. Lessons drawn from these and other readers' experience and assessment of the value of the Reading Tools will assist our understanding of the nature of online reading, the potential readership of online research, the role of context in reading, while the study is also intended to contribute to improving the design of journals and other informational resources in online environments.

References

Siemens, R. G. (2001). "*Unediting and Non-Editions*." In The Theory (and Politics) of Editing. Anglia 119.3: 423-455. Reprint of "*Shakespearean Apparatus,*" with additional introduction. "*Shakespearean Apparatus? Explicit Textual Structures and the Implicit Navigation of Accumulated Knowledge*." Text: An Interdisciplinary Annual of Textual Studies 14. Ann Arbor: U Michigan P, 2002. 209-240. Electronic pre-print published in Surfaces 8 (1999): 106.1-34. < http://www.pum.umontreal.ca/revues/surfaces/vol8/siemens.pdf>.

Siemens, R. G. (2004). "*Modelling Humanistic Activity in the Electronic Scholarly Edition*." Presented at The Face of Text (CaSTA: The Third Canadian Symposium on Text Analysis Research), McMaster U (November).

Siemens, R.G., Susan Schreibman, and John Unsworth. (2004). *The Blackwell Companion to Digital Humanities*. Oxford: Blackwell. xxvii+611 pp.

Siemens, R. G, et al. (2002). The Credibility of Electronic Publishing: *A Report to the Humanities and Social Sciences Federation of Canada*. R.G. Siemens (Project Co-ordinator), Michael Best and Elizabeth Grove-White, Alan Burk, James Kerr and Andy Pope, Jean-Claude Guédon, Geoffrey Rockwell, and Lynne Siemens. Text Technology 11.1: 1-128. < http://www.mala.bc.ca/~siemensr/hssfc/index.htm>.

Siemens, R. G, and Jean-Claude Guédon. (forthcoming). "Peer Review for Humanities Computing Software Tools: *What We Can Draw from Electronic Academic Publication*." A section of "*Peer Review for Humanities Computing Software Tools*" an article cluster, in progress, with Stéfan Sinclair, Geoffrey Rockwell, John Bradley, and Stephen Ramsay.

Warwick, Claire. "*No Such Thing as Humanities Computing?*" Presented at ALLC/ACH, Tuebingen, 2004 <http://www.hum.gu.se/allcach2004/AP/html/prop8.html>.

Willinsky, J. (in press). *The case for open access to research and scholarship*. Cambridge, MA: MIT Press.

Willinsky, J. (2003). *Policymakers' use of online academic research*. Education Policy Analysis Archives, 11(2). Retrieved October 6, 2003, from http://epaa.asu.edu/epaa/v11n2/.

Willinsky, J. (2002). Education and democracy: The missing link may be ours. Harvard Educational Review, 72(3), 367-392.

Willinsky, J. (2001). *Extending the prospects of evidence-based education*. IN>>SIGHT, 1(1), 23-41.

Willinsky, J. (2000). If only we knew: *Increasing the public value of social science research*. New York: Routledge.

Willinsky, J. (1999). The technologies of knowing: *A proposal for the human sciences*. Boston: Beacon.

Willinsky, J., and Forssman, V. (2000). A tale of two cultures and a technology: *A/musical politics of curriculum in four acts*. In C. Cornbleth (Ed.), Curriculum, politics, policy: Cases in context (pp. 21-48). Albany, NY: State University of New York Press.

Music and Meaning in a Hopkins "*Terrible Sonnet*"

Stephanie SMOLINSKY

Humanities Department,
New York City Technical College, CUNY

In this paper, I will examine the phonetic/phonological patterning of a sonnet by Gerard Manley Hopkins, *No Worst, There Is None*, one of the sequence of 'Terrible Sonnets' written during a period of deep despair toward the end of his life. My paper illustrates the use of a computer program, the Pattern-Finder, to develop the conventional, close-reading techniques in literary criticism, as exemplified in Burke (1973) or Vendler (1988), beyond what can be discovered by even the most aware naked eye and ear. I will argue that by using this new computational approach, we can discover a level of meaningful sound-structure present in the poem but previously inaccessible. (A full description of the computationally-based technique which I will be using is found Smolinsky & Sokoloff's abstract proposal for the poster presentation plus software display *Introducing the Pattern-Finder* (#140), immediately preceding this abstract.)

Hopkins was an accomplished linguist (in the sense of 'language-learner:' see Davies 1998), and one of the most consciously linguistically aware of all poets writing in English. He truly knew his medium, and used it to the fullest. In *No Worst*, we see this in his syntax, as in the Shakespearian epithet "no-man-fathomed" or the elided "that" relative in "under a comfort [that] serves in a whirlwind." Both compressions embody the excruciating pressure which the poet is expressing; the second one also gives a sense of uncontrollable velocity.

We see Hopkins' awareness of his medium just as much in his phonetics. One example is the striking /iy/ assonance in six of the end-rhymes, also echoed inside the lines, in "heave," "leave," "shrieked," "steep or deep" and "each," /iy/ being the vowel with the highest pitch (Peterson and Barney 1952): in this context, I would argue, the most evocative of a scream. Another is the /p/ alliteration in the first two lines, all in *pre-stressed* ('opening' (de Saussure 1959), maximally plosive, aspirated (Ladefoged 1982) and forceful) position:

"pitched past pitch" "pangs" "(fore)pangs" the complete absence of /p/ in lines 3-9, and the gradual accumulation of /p/'s in final ('closing' de Saussure op.cit, non-aspirated (Ladefoged op cit), minimally forceful position in lines 10-14. The movement of the poem is from increasing paroxysms of pain, up to a climax and then down to an uneasy annihilation in "sleep," the last word of the sonnet. The placing of the two groups of /p/ alliterations, the strong opening /p/'s at the beginning and the weak closing /p/'s at the end, embody this progression.

So far, I have given instances of Hopkins' control of his medium, the English language, which anyone willing to read slowly and carefully can find for him- or herself. But I would argue that his capacity to build meaningful linguistic structures in his poem goes well beyond what even the most conscientious reader can pick out. If we look below the level of normal speech-sound repetitions (rhyme, alliteration, assonance) we will see that patternings of phonetic/phonological features of the speech sounds in the sonnet also support, even embody, its meaning. Aided by the Pattern-Finder, I will give two simple examples.

The first is the distribution of pre-tonic voiceless consonants, that is to say, those consonants in the positions spotlighted by stress (see *illustration 1*: file Hopkins C cons-stress voiceless). We can think of voiceless consonants (speech sounds which are in the minority among English consonants, and in an even smaller minority if we consider the total number of English speech sounds) as a small—and especially salient *because* small—number of interruptions to the stream of vocal vibration ocurring as the poet or reader recites the sonnet. These interruptions will take the form either of small explosions (plosives) or small frications (fricatives) or something of both (affricates). We see how noticeable they are in the explosive first line's "**p**itched **p**ast **p**itch of grief," or the dragging, elongated quality of the tenth line's "**f**rightful, **sh**eer no-man-**f**athomed! **H**old them **ch**eap"

There are other things to be said about the function of pretonic voiceless consonants in the sonnet (see, again, *illustration 1*: file Hopkins C cons-stress voiceless), but here we will focus only on distributional gaps—where they are absent. We note that while there may be as many as five in a line (see frequency counts (leftmost, in red) for lines 2, 5 and 10), there is only one line with a single pre-tonic voiceless consonant, and only one with none at all. Line 4, with no pre-tonic voiceless consonants, is one in which the poet appeals to the Virgin Mary for "relief." The forcefulness of his expression of distress is softened by the maternal presence which he is addressing; he begs and wails as a small child would. The last line, 14 evokes the dropping away of painful sensation into oblivion; the paroxysms are temporarily quieted. The final thing the poet focuses on is the state of "sleep," and this is the only word in the line picked out by an initial voiceless consonant.

Another feature-distribution worth mentioning, this one a contrast, is that of the stressed front versus the stressed back vowels in the sonnet. Stressed front vowels very much predominate: they are just over twice as frequent as stressed back vowels (see *illustration 2*: file Hopkins V stress front vs stress back). In the literature on phonetic symbolism, an area in which much work has been done on symbolic attributions to speech sounds, such as brightness and darkness, largeness and smallness, front vowels have been found to be bright, active, small, sharp and fast, and back vowels, dark, passive, large, dull and slow (see Sapir 1929, Newman 1933, Grammont 1946 among many others). So, given that the sonnet is about suffering, our finding might at first seem counterintuitive. But then we consider that the suffering described is acute ("Pitched past pitch of grief"), and that the consciousness evoked in the poem is just as intensely pleading for release ("my cries heave"). Moreover, the form of release grudgingly granted ("Here! Creep/ Wretch") is only the temporary one of sleep. Thus, we realize that the sharp, bright, wakeful quality of the front vowels is apt for the evocation of a painful emotional state which is bound to recurr. We see that in line 11 and the beginning of 12, where the largest cluster of back vowels in the poem is found, these dominate for only a minute stretch: "N**or** l**o**ng does **ou**r sm**a**ll d**u**rance…" but are then taken over by the front vowels again "d**ea**l with that st**ee**p or d**ee**p. H**ere**! Cr**ee**p/Wr**e**tch.." Finally, the last line is the only one without any back vowels: there will be no true escape into something larger than the suffering self. The consciousness is to be granted only a temporary respite in the "whirlwind."

This is a preliminary sample of the kind of work one is able to do using the Pattern-Finder. As I said earlier, Hopkins was an accomplished linguist; he is known for two terms, 'inscape' and 'instress' that suggest the influence of subliminal awareness on poetic meaning. His concepts would lead us to believe that his poetry is a natural candidate for our approach:" The word 'inscape'

(coined by analogy with 'landscape') varies in its implications. But its main meaning is distinctive pattern, the relationship between parts that creates the integrity of the whole, which in turn is different at different times. 'All the world is full of inscape and *chance left free to act falls into an order as well as purpose*.' So there is pattern even in natural accident…No two things, if properly seen, are identical. Individuality is irreplaceable. The key-note of inscape is therefore not just pattern, but *unique* pattern." (Myitalics) "Instress is the active energy that binds parts of the inscape of the whole…It is also a faculty of the human mind when it brings things into creative relationship. It demands an act of pure attention…for 'the eye and the ear are for the most part shut and instress cannot come…' " (Davies, 1998, quoting Hopkins)

References

Burke, K. (1973) On Musicality in Verse in *The Philosophy of Literary Form*, Berkely, CA: University of California Press

Davies, W, (ed) (1998) *Poetry and Prose of Gerard Manley Hopkins,* London, UK: J. M. Dent

Grammont, M. (1946) *Traité de Phonétique*, Paris: Delagrave

Ladefoged, P. (1982) *A Course in Phonetics*, New York, NY: Harcourt, Brace Jovanovich Inc.

Martin, R. B. (1991) *Gerard Manley Hopkins: A Very Private Life*, London, UK: HarperCollins

Newman, S. (1933) Further experiments in phonetic symbolism, *American Journal of Psychology* v 45, pp 53-75

Peterson, G.E. and H. L. Barney, (1952) Control methods used in the study of the identification of vowels, *Journal of the Acoustical Society of America* v 24 # 2, pp175-184

Sapir, E. (1929) A study in phonetic symbolism, *Journal of experimental psychology*, v 12, pp 225-239

Saussure, F. de (1959) *Course in General Linguistics*, New York, NY: The Philosophical Library

Vendler, H (1988) *The Music of What Happens,* Cambridge, MA: Harvard University Press

White Rabbit: A Vertical Solution for Standoff Markup Encoding and Web-Delivery

Carl STAHMER

MITH, University of Maryland

Below is an extract of a conventional, TEI encoded stanza from an Early Modern ballad by an anonymous author:*

> For what we with our flayes coulde get
> To kepe our house and seriauntes
> That dyd the freers from us fet
> And with our soules played the marchantes
> And thus they with theyr false warantes
> Of our sweate have easelye lyved
> That for fatnesse theyre belyes pantes
> So greatlye have they us deceaved

This form of XML markup, which imbeds the markup within the artifact itself, will be familiar to most readers. In their 1997 essay, "Hyperlink Semantics for Standoff Markup of Read-Only Documents", Henry S. Thompson and David McKelvie describe a system of "standoff markup" in which markup data resides in a physically separate file from that which it describes. Using a standoff markup system of the type described by Thompson and McKelvie, all of the TEI markup from the above citation would be stored in a separate file from the text itself, relying on a pointer system to describe its relation to the root text:*

The above example uses the order of words in the root text to define inclusionary collections of words that are described by the XML elements in the standoff markup file. It provides the same markup description of the text as the more familiar, traditional example given above, but without intruding into the integrity of the root text file.

As noted by Thompson and McKelvie, one major advantage of a standoff markup system is that it presents the condition of possibility for utilizing multiple, overlapping hierarchies to describe the same root artifact—something that cannot be achieved using a conventional, single-file

markup approach. At the Maryland Institute for Technology and the Humanities (MITH, http://www.mith.umd.edu), we are currently collaborating with the Library of Congress (LOC, http://www.loc.gov/) to develop a platform for the encoding and delivery of digital resources in the LOC's American Memory collection (http://memory.loc.gov/ammem/) utilizing an interactive, standoff markup system designed specifically to allow multiple, overlapping markup hierarchies, including markup provided by web-users visiting the collection via the American Memory website.

This platform, named White Rabbit, is a vertical standoff markup solution that provides an easy to use Graphical User Interface (GUI) for editorially controlled base-document preparation, a collection of web-service applications that allow users to browse, search, and retrieve recourses using a standard, HTML web-browser or to retrieve raw XML source for each resource, and, most importantly, provides a web-based interface for users to add their own markup layers to texts in the collection.

On the technical side, White Rabbit functions by tokenizing raw, ascii textual data at the level of base, lexically significant units (most often words, but frequently other diacritical and textual elements) and storing an ordered list of tokenized elements in a SQL database. It then allows users to construct XML using a simple point and click interface and to validate this XML against a DTD. XML "layers," including those created by resource consumers visiting the LOCs website, are stored in a collection of related database tables.

As new markup layers are added to each artifact, resource consumers gain the ability to choose which markup "layers" to apply to the text on delivery. Once a markup layer is chosen, users can then perform advanced XML searching, parsing, and manipulation of artifacts using any web browser. For example, using the example of the Early Modern ballad extracted above, the user could search within a single or collection of Ballads for all instances of the occurrence of a particular word or phrase within a refrain only. Using convential markup systems, this type of functionality is available only using specialized XML browsers.

White Rabbit also provides a collection of "hierarchy analysis" tools that allow users to analyze the ways in which multiple markup layers for a given text relate to each other. Using these views, a user can identify filtered or un-filtered collections of layers and search for statistically significant patterns of convergence and divergence between multiple markup layers in these collections. Over time, this aspect of White Rabbit's functionality provides an increasingly valuable bank of data regarding the ways in which a growing collection of users understand both formal and thematic elements of artifacts contained in the collection.

With White Rabbit you can have your cake and eat it too, applying multiple markup strategies to the same text for retrieval and display when determined by particular scholarly contexts, providing robust analysis of the patterns in textual structure that emerge through multiple, overlapping markup layers, and delivering finely tuned XML parsing and searching of texts to any user with a standard web-browser.

White Rabbit is comprised of a collection of cross-platform, Java-based client and server applications and applets that communicate with a SQL database. It is an open-source platform specifically designed to be easily exportable to a variety of platforms and uses and will be available for public download in both compiled and source distributions in late 2006.

The proposed paper will present a brief introduction to the concept of standoff markup as described above, an interactive demonstration of the advantages of standoff markup, and, finally, an interactive demonstration of the White Rabbit platform. Detailed information about how to download and implement White Rabbit and/or participate in open-source project development will also be provided.

* please note that submission of XML examples above in plain-text was breaking the automated the submission interface. As such, entity references were used to replace certain elements. This will dramatially decrease the readability of example if abstract is read in plain-text.

A Mathematical Explanation of Burrows's Delta

Sterling STEIN
stein at ir.iit.edu

Shlomo ARGAMON
argamon at iit.edu

Linguistic Cognition Laboratory
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616

Introduction

While many methods have been applied to the problem of automated authorship attribution, John F. Burrows's "Delta Method" [1] is a particularly simple, yet effective, one [2,3] The goal is to automatically determine, based on a set of known training documents labeled by their authors, who the most likely author is for an unlabeled test document. The Delta method uses the most frequent words in the training corpus as the features that it uses to make these judgments. The Delta measure is defined as:

> the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text [4].

The Delta of the test document is computed with respect to each of the training documents, and that author whose training document has minimal Delta with the test document is chosen for attribution.

While this method is intuitively reasonable, we may still ask: Why does this method work so well? Why z-scores? Why mean of absolute differences? Perhaps if we understood the mathematical underpinnings of the method, we could modify it to make it more effective. Furthermore, we would know better when it is applicable and when using it would not make sense.

Hoover has implemented Delta-based attribution as a spreadsheet [5]; we have re-implemented it as Java program with several options for using different variations, which we are experimenting with. This program will be made available to the research community once it has reached a stable state.

Probabilistic formulation

This section will briefly show how Burrows's Delta may be profitably viewed as a method for ranking authorship candidates by their probability. Let X and Y be n-dimensional vectors of the word frequencies in two documents. Note that the z-score is obtained by subtracting out the mean and dividing out the standard deviation. Then the Delta measure between these documents can be reformulated, as shown in Figure 1.

Note that the value of mean[i], the mean frequency of word i, cancels out, with the only effect of the training corpus as a whole being the normalizing factor of std[i], the standard deviation for word i.

Thus, Delta is like a scaled distance between the 2 documents. It is not the ordinary distance "as the crow flies", but rather it is the sum of each dimension independently, called the "Manhattan distance". It is like walking the streets of Manhattan as we stay on the grid.

Note that if we consider the mean of a distribution in place of Y[i], this has a form similar to a Laplace probability distribution [6]. Specifically, it is the exponent of the product of independent Laplace distributions. Thus, we are assuming that the individual document that we are comparing the testing document against is a sort of average document for that author. Taking of the z-score corresponds to the normalization in the exponent. So, in a sense, Delta is measuring the probability of a document being written by an author taking each word frequency independently and then choosing the document with the highest probability.

In effect, that we are using the z-score means that we are estimating the parameters of the Laplace distribution by the sample mean and standard deviation. However, the maximum likelihood estimator of the Laplace distribution is the median and the mean absolute deviation from the median [6]. This gives us our first variation of the Delta measure. Instead of using the z-score, we should use the median and "median deviation". Whereas

the Delta measure gives a distance in a purely abstract space, this variation provides a well-founded probability.

Now that we know we are looking at a probabilistic model, we can try putting in other distributions. A commonly-used distribution is the Gaussian, or normal, distribution. It is similar to the Laplace distribution except that it uses a sum of squares rather than of absolute values, based on the mean of the mean and standard deviation, hence using the z-score is appropriate here. Note that in this case, we have the "as the crow flies" Euclidean distance instead of the Manhattan distance.

Further, note that the previous measures consider each dimension independently. In this sense, they are axis-aligned. This means that the use of each word is assumed to have nothing to do with the use of any other words. Of course, this assumption is false, but may be a reasonable approximation. To take this co-occurrence into account, we can use a rotated method, eigenvalue decomposition. Previously we used the z-score of individual words. Instead of using the standard deviation, we can generalize to using the entire covariance matrix. In this, we take the largest magnitude eigenvalues from the covariance matrix and use the corresponding eigenvectors as the features.

Evaluation

We are currently performing empirical tests. To compare these new variants to the original and each other, we will use 3 corpora. First, we will use the data in the spreadsheets from [5] to check that our implementation is working properly and so we can directly compare results. The second will be a collection of essays written by students taking a psychology course. There are up to 4 essays by each author. The third will be the 20 newsgroups corpus from http://people.csail.mit.edu/jrennie/20Newsgroups/. These corpora will be split into testing and training such that the training has only 1 or 0 of each author. Having 0 allows for the possibility of an unknown author. Each of these variations will be run on the corpora and the results of each classification will be split into 5 categories, based on who the attributed author is (lowest Delta candidate), and whether the attribution is considered reliable. This is determined via a threshold on attribution confidence, such that confidences below the threshold are considered "unknown". To determine the confidence, we use Delta-z, following Hoover [3].

Say/Is	A	B	U
A	a	b	g
B	b	a	g
U	d	d	e

The first type of classification decision, a, is a true positive, where the correct author is attributed. Next, b is a false positive, where a known author is chosen, but not the correct one. Another false positive is g, where a known author is chosen, but the true author is not in the training. Fourth, d is a false negative, where no author was recognized, but should have been. Finally, e is a true negative, where the true author was not in the training set and was not recognized.

These allow us to calculate the true positive and false positive rate, where:

true positive rate = a/(a+d)

and

false positive rate = (b+g)/(b+g+e)

These values can be used to make a receiver operating characteristic (ROC) graph (Figure 2), which shows the sensitivity of a method to the signal in the data. It is trivial to declare everything a negative, which would give 0 to both TP and FP. As you classify more instances as positive, there is more of a risk that an instance classified as positive is not. An overall measure of a method's efficacy can be computed as the area under the ROC curve (AUC); this score will be used for comparison between the methods. AUC measures the trade off between false positives and false negatives, with the baseline at 50% (where the line goes straight from 0,0 to 1,1) and the best possible value of 1, meaning always getting true positives with no false positives. In this way, it will allow us to judge and compare how well the different variations work.

Conclusion

We have reformulated Burrows's Delta method in terms of probability distributions. This allows us to extend the method to use multiple different probability distributions and to interpret the result as a probability. At the conference, we will present results comparing the effectiveness of these variations. More importantly, this

work provides a more solid foundation for understanding Delta. In particular, the probabilistic assumptions that it makes, such as word frequency independence and that authors have similarly-shaped word-frequency distributions, are made explicit, allowing us better understanding of the uses and limitations of the method. For example, it is now clear why the method should only be applied to documents all of the same well-defined text type.

$$\begin{aligned}\sum_{i=1}^{n} |z(X_i) - z(Y_i)| &= \sum_{i=1}^{n} \left| \frac{X_i-\mu_i}{\sigma_i} - \frac{Y_i-\mu_i}{\sigma_i} \right| \\ &= \sum_{i=1}^{n} \left| \frac{X_i-\mu_i-(Y_i-\mu_i)}{\sigma_i} \right| \\ &= \sum_{i=1}^{n} \left| \frac{X_i-Y_i}{\sigma_i} \right|\end{aligned}$$

Figure 1: Reformulating Burrows's Delta as a distance measure.

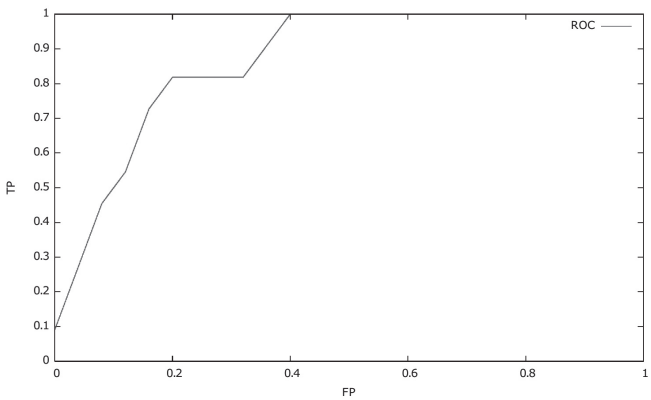


Figure 2: A sample ROC curve. The X-axis is the false positive rate FP, while the Y-axis is the true positive rate TP. Note that TP increases monotonically with FP, and that TP=1 once FP=0.4.

References

[1] **Burrows, J. F.**. (2002a). *Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship*. Literary and Linguistic Computing 17, pp. 267--287.

[2] **Hoover, D.**. (2004b). *Delta Prime?*. Literary and Linguistic Computing 19.4, pp. 477--495.

[3] **Hoover, D.**. (2004a). *Testing Burrows's Delta*. Literary and Linguistic Computing 19.4, pp. 453--475.

[4] **Burrows, J. F.**. (2002). *The Englishing of Juvenal: Computational Stylistics and Translated Texts*. Style 36, pp. 677--699.

[5] **Hoover, D.**. (2005). *The Delta Spreadsheet*. Literary and Linguistic Computing 20.

[6] **Higgins, J., Keller-McNulty, S.**. (1994). *Concepts in Probability and Stochastic Modeling*. Duxbury Press, 1 ed..

[7] **Juola, P.**. (2005). *A Prototype for Authorship Attribution Software*. Literary and Linguistic Computing 20.

Annotation en mode collaboratif au service de l'étude de documents anciens dans un contexte numérique

Ana STULIC
AMERIBER, Bordeaux 3 University

Soufiane ROUISSI
CEMIC - GRESIC, Bordeaux 3 University

Les mutations liées au numérique, nous invitent à revisiter une approche traditionnelle d'édition de textes telle qu'elle est pratiquée dans les sciences humaines en tant qu'outil de recherche. Notre proposition s'applique au problème d'édition électronique des documents judéo-espagnols en écriture hébraïque sur lesquels des techniques de translittération et transcription sont mises en oeuvre. Nous étudions les possibilités de construction d'une plate-forme collaborative sur laquelle les documents pourront être annotés et organisés en tant que réservoir numérique de référence au service des chercheurs de ce domaine spécifique. Au delà de la description nécessaire à l'identification des documents (la norme descriptive Dublin Core), notre objectif est de favoriser l'annotation (sous une forme libre, complémentaire à la méta-description) des documents enregistrés à des fins d'étude et de discussion. Les documents sources (sous forme de fichiers images, par exemple) sont transformés par la transcription or translitteration en documents numériques qui deviennent des outils de recherche. L'annotation (effectuée par le chercheur ou expert du domaine) intervient à la fois pour métadécrire la source et les documents " dérivés " (à des fins d'identification, de localisation des ressources appartenant au réservoir numérique) mais aussi pour permettre une mise en discussion (interpréter, commenter, réfuter, traduire, accepter …).

Un premier travail nous a permis de proposer un modèle conceptuel définissant la structuration des données nécessaire à ce double objectif d'identification et d'annotation (Rouissi, Stulic 2005).

Nous adoptons le point de vue selon lequel la mise en place réussie d'un dispositif socio-technique adapté aux besoins des chercheurs d'une communauté scientifique dépend de la définition pertinente de ces fonctionnalités. Pour les définir dans ce contexte précis, qui concerne une communauté scientifique relativement réduite et géographiquement dispersée, nous avons prévu la réalisation d'une enquête auprès des chercheurs concernés. Cette enquête a pour objectif de confronter nos hypothèses initiales (modèle conceptuel de définition des annotations) avec les pratiques déclarées par les universitaires. Les résultats de l'enquête doivent nous permettre d'identifier les besoins et les attentes des chercheurs interrogés de façon à déterminer des pistes de développement et de mise en place de fonctionnalités adaptées. Il s'agit également de mesurer les niveaux d'intégration des technologies de l'information et de la communication (TIC) dans leurs pratiques actuelles (types d'outils technologiques utilisés, cadre individuel ou collectif, recours à l'annotation sur des documents numériques, participation éventuelles à des environnements collaboratifs …). L'analyse des réponses se fera à un niveau quantitatif (questions fermées) et à un niveau qualitatif (analyse de contenu) pour nous permettre de proposer une typologie des pratiques déclarées et/ou des besoins.

Dans les lignes générales, notre travail s'oriente vers l'élaboration d'une plate-forme accessible en mode full-web et basée sur des solutions existantes respectant notamment les principes de l'Open Source. Mais ce sont les fonctionnalités attendues par les membres de la communauté qui vont modeler les choix techniques concrets.

Nous envisageons un format propriétaire pour le modèle en question, mais pour des raisons de portabilité celui-ci devra permettre des exportations basées sur des formats d'échanges de données. Ces formats doivent s'appuyer sur des schémas largement partagés et respectant une démarche normative comme la Text Encoding Initiative (TEI) pour la structuration des documents ou encore le Dublin Core pour la description des documents de travail.

La plate-forme devra favoriser l'annotation en mode collaboratif (avec divers niveaux possibles de participation : du lecteur à l'éditeur) et organiser la discussion (celle-ci ayant plusieurs buts possibles : de transcrire, de traduire, de commenter, de proposer …) tout en la caractérisant (selon des vocabulaires restant à construire, pas forcément pris en compte dans les spécifications comme la TEI).

Le dispositif final comprendra aussi la construction d'un corpus documentaire numérique qui sera exploitable grâce à des fonctionnalités de recherche (dans un but d'identification, de sélection…) des documents. Un futur utilisateur du dispositif pourra dans un premier temps effectuer une recherche (de type sélection) sur les documents numériques constituant le corpus, puis après sélection d'un document, consulter la discussion sur celui-ci et selon son rôle (son profil) il pourra éventuellement participer en annotant le document et/ou en réagissant à une des annotations existantes. L'utilisateur pourra générer une nouvelle version comportant ses propres annotations et exporter celle-ci selon les formats et standards (à prévoir et à implémenter) qui lui seront proposés. Malgré l'apparente singularité de notre contexte lié au traitement des textes judéo-espagnols, notre réflexion est menée vers une approche plus large. A terme, notre but est de construire un environnement numérique qui pourrait être appliqué dans les domaines de recherche avec des besoins similaires autour de l'organisation de corpus documentaires à des fins d'étude.

Annotations in Collaborative Mode for Ancient Documents Study in Digital Environment

The development of digital technology make us revisit the traditionnal approach of text edition as a research tool such as used in the Humanities. Our proposal applies to the problem of electronic editing of Judeo-Spanish documents in Hebrew characters over which the translitteration and transcription techniques are carried out. We study the possibilities of collaborative platform construction in which these documents could be annotated and organized as a reference digital repository for the scholars of this specific domain. Beyond the meta description necessary for the document identification (Dublin Core metadata standard), we aim to promote the annotating of the uploaded documents (in the free form, complementary to the metadescription) for the purpose of study and discussion. The source documents (presented as image files, for instance) are transformed into digital documents, which become then the research tools. The role of annotation (carried out by a researcher or a domaine expert) is to convey the metadata information about the source and the " derived " documents (for the purpose of identification and localization of digital repository resources) and to make possible the discussion over these documents (to interprete, comment, refuse, accept, translate, and so forth). In a previous paper we proposed a conceptual model defining the data structuring necessary to this double purpose of identification and annotation (Rouissi, Stulic 2005).

Our point of view is that the successful setting up of socio-technical device for the researchers of a particular scientific field depends on the definition of appropriate functionalities. In order to define them in this context of relatively small and geographically dispersed scientific community, we planned to realise a survey that should confront our initial hypothesis (our conceptual annotation model) with the declared academic practices. The survey results should allow us to identify the needs and the expectations of researchers and to determine the direction of project development and the setting up of appropriate functionalities. It should also measure the level of Information and Communication Technology (ICT) integration in their current activities (the type of technological tools in use, the individual or collective framework, the use of annotation over digital documents, the participation in collaborative work environments…). The answers will be analysed at the quantitative (closed questions) and the qualitative level (the content analysis) so as to provide a typology of declared practices and/or needs.

In broad lines, our work is oriented towards the elaborating of the full Web platform and is based on the existent open source solutions. But only with the clear definition of the functionalities that are expected by the scientific community members we will be able to model the final technical choices.

We envisage the proprietary format for the model in question, but it should permit the exportation through the standardized formats for data interchange like Text Encoding Initiative for document structuring or Dublin Core for meta description of working documents. The platform should promote the annotating in collaborative mode (with different levels of participation : from reader

to editor) and organize the discussion (which can can different purposes : to transcribe, to translate, to comment, to propose…) by qualifiying it (following the vocabularies that are still to be defined, not necessarily taken into account in the specifications such as TEI). The final platform will include also the documentary corpus building that will be exploitable thanks to the search functionalities.

The future user will be able to effectuate a search (of 'selection' type) among digital documents which constitute the corpus, and after the document selection, to consult the related discussion and according to his/her profile, he/she will be able to participate by adding the annotation and/or reacting to the annotation previously added by other user(s). The user will be able to generate a new version with his/her own annotations and to export it following the proposed formats and standards.

Although the context to which we apply our project is very specific, our reflexion is oriented towards a larger approach. Our aim is to construct a digital environment which could be applied to the research domains with similar needs for documentary corpus organisation and work on documents.

References

Rouissi, S. Stulic, A. (2005). *Annotation of Documents for Electronic Edition of Judeo-Spanish Texts: Problems and Solutions*, à paraître dans les Actes de la conférence Lesser Used Languages and Computer Linguistics, Bolzano, EURAC, 27-28 octobre 2005. Résumé consultable en ligne : http://www.eurac.edu/NR/rdonlyres/9F93F5B9-95F6-44AC-806D-C58FF69AFD27/8812/ConferenceProgramme2.pdf

TEI P5, Guidelines for Electronic Text Encoding and Interchange, C.M. Sperberg-McQueen and Lou Burnard, 2005, Text Encoding Initiative Consortium, http://www.tei-c.org/

Strings, Texts and Meaning

Manfred THALLER

Universität zu Köln

From a technical point of view, texts are represented in computer systems currently as linear strings of atomic characters, between which no distinction is being made on the technical level. In the markup discussions within the Humanities, this is usually accepted as an immutable fact of technology.

We propose, that the handling of Humanities texts could be considerably easier, if an engineering model could be created, which is built upon a more complex understanding of text.

1. Basic model of "text" proposed

For this we start with the proposal to understand a text as a string of codes, each of which represents "meaning" measurable in a number of ways.

More detailed:

Texts – be they cuneiform, hand written or printed – consist of information carrying tokens. This tokens fall into a number of categories, which are differentiated by the degrees of certainty with which they can be used in various operations. The trivial example are ASCII or Unicode characters. Less trivial are symbolic tokens, like, e.g. the (primitive) string representing the term "chrismon", a bit map representing a Chrismon (or something similar) etc.

A string made up of such tokens, which represents a text, can be understood to exist in an n-dimensional conceptual universe. Such dimensions, which have different metrics are, e.g.:

- A dimension which has coordinates with only two possible values ("yes", "no") which describes, whether a token has an additional visible property, like being underscored.
- Another dimension, which has coordinates on a metric scale, which assigns a colour value, which allows to define similarities.
- Another dimension describing the position of a

token like "Chrismon" with a ontology describing the relationships between Chrismons and other formulaic forms.

- A real number, giving the relative closeness between a bitmap representing a Chrismon and an idealtypische Chrismon.

If we view such a string from a specific point in the conceptual space – a.k.a. an individual's research position – many of these dimensions tend to collapse in the same way, as 3 dimensional objects collapse their z-value when represented in two dimensional drawings.

2. Relationship between text, markup and processing

We assume, that string processing, on a very low level of engineering, can be implemented in such a way, that the low level programming tools, which are used today for the generation of programs handle texts, can tackle the implications of this model directly. This implies, e.g., a low level function, which can compare two strings "sensitive for differences between included symbolic tokens beyond a specified ontological distance" or "insensitive for this" very much like current implementations of low level tools can compare can compare two strings as "case sensitive" or "case insensitive".

While currently all textual phenomena have to be described with one integrated system of markup, expressing attributes, which can only be observed on the character level, without necessarily being interpretable on the spot, as well as highly abstract textual structures, the proposed approach would divide textual attributes into two classes: *Textual* attributes in the more narrow sense, which can be handled as properties of the strings used to represent the texts and *structural* (and other attributes) which are handled by a software system implying the presence of the underlying capabilities of the low level textual model, while focusing itself upon a class of higher level problems: E.g. a data base operating upon an abstract content model of a drama, relying upon the handling of page references as well as critical apparatus by the underlying string handling tools.

The later implies that documents will – seen from today's perspective – usually be marked up in at least two concurrent ways. Some implications of that will be listed.

3. Possibilities of generalizing the basic model.

Our model so far has assumed, that information is handled by strings, i.e. by tokens which form one-dimensional sequences. (Non linear structures are one-dimensional as well in this sense: a path within a graph has a length, measured as the number of nodes through which it passes. It cannot be measured in two dimensions, as the relative location of the nodes within a two dimensional drawing is just a property of the visualization, not the structure itself.)

There is no reason, however, why the notion of meaning represented by an arrangement of tokens carrying information should not be generalized to two dimensions (images), three dimensions (3D objects) or four dimensions (e.g. 3D representations of historical buildings over time).

A problem arises, however, when one compares some operations on one- with the same operations on more-dimensional arrangements of information carrying tokens. A good example is the comparison of "insertion operations" in strings v. the same operation in images.

We conclude by proposing to solve that problem by the notion, that a textual string is a representation of an underlying meaning with a specific information density, which usually will transfer only part of the meaning originally available, just as a digital image represents only part of the visual information available in the original.

This in turn leads to the notion, that not only the handling of information carrying tokens can be generalized from the one to the more-dimensional case, but the properties of markup languages can as well.

4. Concluding remark

While the generalisation of the model quoted above is presented in Paris for the first time, the idea of a specilised data type for the representation of Humanities text goes back to the early nineties (Thaller 1992, Thaller 1993). Various intermediate work never has been published, an experimental implementation, focusing on the interaction between texts and databases

administering the structure embedded into the text does exist, however and is used in the production level system accessible via http://www.ceec.uni-koeln.de (Thaller 2004). More recently a project started at the chair of the author, to implement a datatype "extended string" as a series of MA theses in Humanities Computer Science. The first of these (Neumann 2006) provides a core implementation of the most basic concepts as a class augmenting Qt and fully integrated into that library.

References

Neumann, J. (2006). *Ein allgemeiner Datentyp für die implizite Bereitstellung komplexer Texteigenschaften in darauf aufbauender Software. Unpubl.* MA thesis, University at Cologne. Accessible via: http://www.hki.uni-koeln.de/studium/MA/index.html

Thaller, M. (1992). "The Processing of Manuscripts", in: Manfred Thaller (Ed.) *Images and Manuscripts in Historical Computing,* Scripta Mercaturae (=Halbgraue Reihe zur Historischen Fachinformatik A 14).

Thaller, M. (1993). "Historical Information Science: Is there such a Thing? New Comments on an Old Idea", in: Tito Orlandi (Ed.): *Seminario discipline umanistiche e informatica. Il problema dell' integrazione* (= Contributi Del Centor Linceo Interdisciplinare 'Beniamo Segre' 87)."

Thaller, M. (2004). "Texts,Databases, Kleio: A Note on the Architecture of Computer Systems for the Humanities", in: Dino Buzzetti, Giuliano Pancaldi, Harold Short (Eds.): *Digital Tools for the History of Ideas* (= Office for Humanities Communication series 17) 2004, 49 - 76.

Modelling a Digital Text Archive for Theatre Studies -- The Viennese Theatre Corpus

Barbara TUMFART

Austrian Academy of Sciences, Vienna

The submitted paper gives an overview of the recently founded project Viennese Theatre Corpus (VTC), a sub-project within the wide range of work undertaken by the Austrian Academy Corpus research group at the Austrian Academy of Sciences in Vienna. In general the presentation will deal with the specific content mark up of theatre-related materials making use of XML and discuss the application of new technologies and the possibilities of advanced research methods for the complex and unique historical and socio-political situation of theatre and its language spoken on stage in Vienna during the 19th century.

Focusing on the Viennese theatre of the 19th century , the playwright Johann Nepomuk Nestroy (1801-1862), is the centre of interest and represents the main starting point for the development of the corpus. Johann Nestroy dominated the commercial stage as actor-dramatist for nearly thirty years in the Austrian theatre of the 19th century and wrote about 80 plays. Using different registers of Viennese German ranging from stagy literary German to colloquial Viennese, the work of this supreme comic dramatist is full of allusion, vivid metaphor and cynical criticism. Emphasising linguistic inventiveness his comedies reflect a deep-rooted discontent with moral double standards, static and mendacious society, false friendship and unfounded prejudices.

The Historical-Critical Edition of the Collected Plays of Nestroy, which appeared between 1924 and 1930, published by Otto Rommel and Fritz Brukner represents the core text of this theatre corpus. With the aim to provide a structured overview of the critical and historical reception of Nestroy's work and a valuable collection of different registers of language used on the stage of the Viennese popular theatre, a large amount of contemporary dramas from others of the 19th century like Friedrich Kaiser, Carl Elmar, Carl Giugno and Alois

Berla will be incorporated. Furthermore historical and socio-economic descriptions, press criticism, texts from the theatre censorship of this period, legal publications and memoirs of famous actors and stage directors offer a well-assorted and wide-ranging collection of different text types.

In its function as a main place for entertainment and social contact, the theatre was expanding in this period. In general the Viennese public was used to about thirty new premieres a year and new plays and new productions in European capitals like London and Paris led to a very rapid production and exchange of theatre texts. As a consequence, translations and adaptations of English and French plays reached the Austrian capital very quickly, often in pirated editions. Therefore Nestroy's work was highly influenced by the literary production of other European countries. Works by Charles Dickens, Eugène Scribe, Paul de Kock, Dupeuty, Bayard as well as Varin served him as stimulation or models for his own plays and for this reason the incorporation of a wide selection of Nestroy's sources from English and French literature is planned.

The challenge for the development of a specific content mark up due to this diversity of text types and language resources is one of the main topics of the paper. The paper will briefly illustrate the annotation system which is used for the Viennese Theatre Corpus, discuss the scope of XML for dramatic texts and illustrate the incorporation of the critical appendix of the printed version of Nestroy's Collected Plays. Due to the contradiction between performed speech, the action on stage and the subsequent publication as printed text, drama is often structurally complicated. The paper will address the dominant question how XML and cognate technologies can help to structure a play, to illustrate the underlying dramatic composition and its use as an analytical tool by means of a specific semantic tagging strategy rather than as a simple archiving solution.

The VTC, a thematically oriented text corpus, will provide access to a variety of original source documents and secondary material. In consideration of the special textual and material requirements of theatrical works the paper will finally discuss the potential contribution of this specific corpus of Viennese plays of the 19th century to scholarly knowledge within the wide range of philological, historical and literary research domains.

Textual Iconography of the Quixote: A Data Model for Extending the Single-Faceted Pictorial Space into a Poly-Faceted Semantic Web

Eduardo URBINA
Hispanic Studies, Texas A&M University

Richard FURUTA
Jie DENG
Neal AUDENAERT
Computer Science, Texas A&M University

Fernando González MORENO
Universidad de Castilla La Mancha

Manas SINGH
Carlos MONROY
Computer Science, Texas A&M University

Since its initiation in 1995, the *Cervantes Project* has focused on creating a comprehensive on-line resource centered on the iconic Hispanic author Miguel de Cervantes (1547-1616). The project, a collaboration of researchers in Hispanic Studies and Computer Science, provides bibliographical, biographical, and textual materials, including an Electronic Variorum Edition based on first editions of Cervantes' best known work, *Don Quixote*. Recent additions include an exhibit of bookplates (*ex libris*) inspired by the *Quixote*. Soon to be released are a significant collection of illustrations associated with key *Quixote* editions (the textual iconography) and a presentation of the musical aspects and influences of Cervantes' works. These latter two collections will be discussed further in this paper.

As our work with the Cervantes Project has evolved, we have become increasingly aware of the need to develop rich interlinkages between the panoply of resources collected for and produced by this project [3]. Within the broader scope of the project as a whole, our textual iconography collection offers a fertile ground for exploring

the needs of large, interdisciplinary collections, and the techniques for integrating the resources they contain. The textual iconography collection (Figure 1) is intended to facilitate a more complete understanding of the iconic transformations of the *Quixote* [5] throughout history by assembling a digital collection of more than 8,000 illustrations taken from over 400 of the most significant editions of this seminal book [15]. In the process of assembling and working with this collection, we have found traditional approaches to presenting image-based collections in both print and digital media to be lacking [12]. While print has long been the dominant media of scholarly communication, it does not scale well to projects of this scope [8]. Digital image collections, on the other hand, are often focused on presenting large numbers of images and associated metadata but lack the scholarly commentary and integration with other textual resources required for scholarly investigations. Moreover, we have found a need to adopt a rich model of hypertext, based on automatically-generated links and multiply-rooted to reflect multiple interpretations of simultaneous significance.

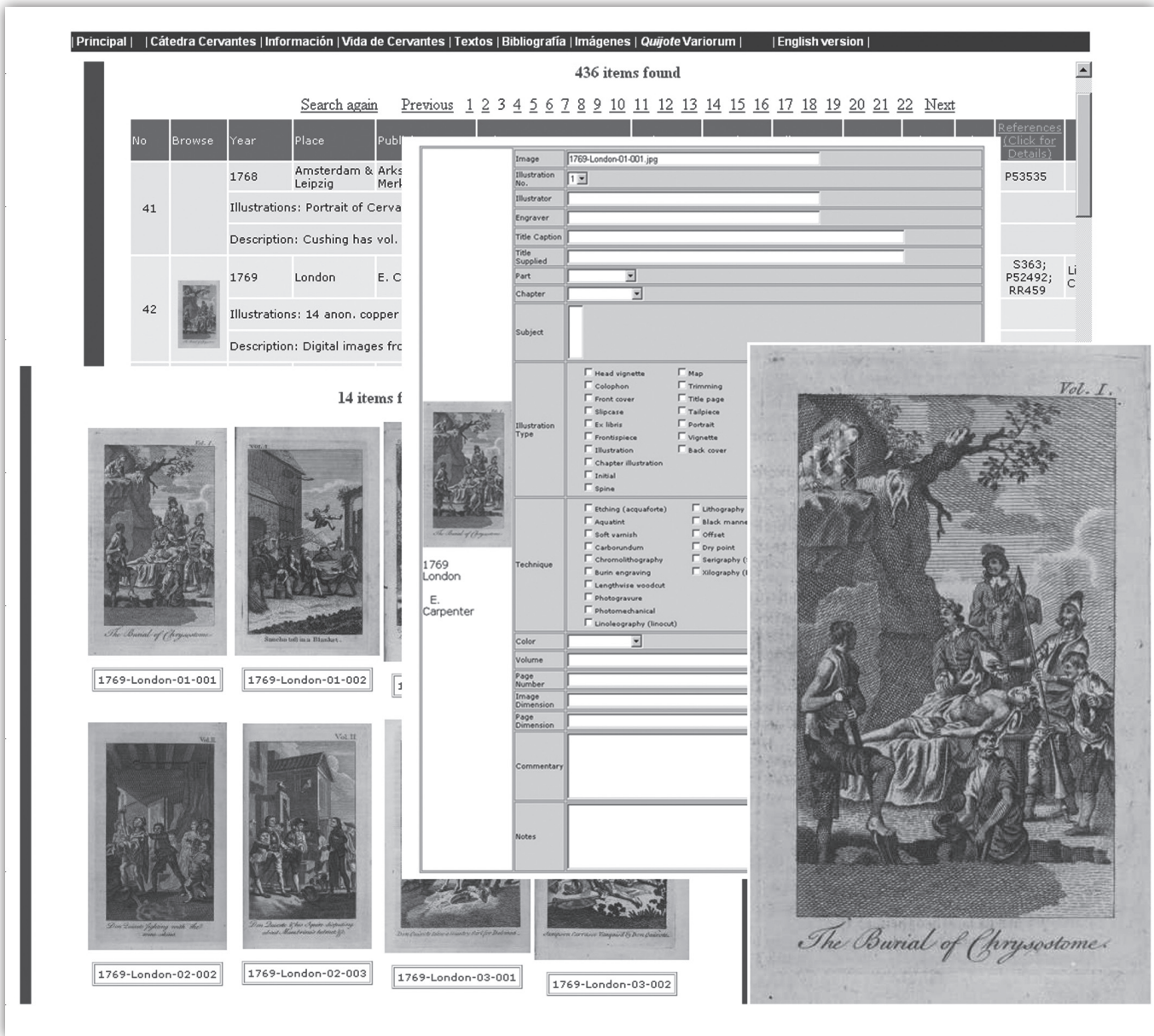


Figure 1: The textual iconography's browsing interface allows viewing of the collection and its metadata. The collection's editors also have access to maintenance tools.

We have aggressively pursued the development of this collection [2] with the intention that it will enable new forms of textual, visual and critical analysis. In particular, it enables scholarly research in two major directions that remain poorly understood despite the incredible amount of scholarly attention devoted to the *Quixote*. First, the textual iconography of the *Quixote* is a key resource to help literary scholars better understand its reception and interpretation throughout history—the illustrations acting as a "hand-mirror [3]" that allows us to see how successive generations of readers, including our own, have literally painted themselves into Cervantes' story [7]. Second, this collection is an invaluable dataset for art historians enabling them to better explore the tools and techniques employed in the often neglected field of book illustration [4]. The illustrations found in the pages of the *Quixote* trace the evolution of graphic art in modern printed works, from the first wood cuts of early 17th century, to the copper engravings, etchings aquatints of the middle and late 17th century and 18th century, to the xilographs and lithographs of the 19th century, and finally the mechanical techniques of the 20th century including offset printing [10,13,14]. These areas remain poorly understood, not simply because of a lack of attention, but in large part because traditional approaches do not adequately support investigations of this nature.

The value of the textual iconography collection is not limited to the scholarly community. Its canonical status in world literature courses, its iconic nature in Hispanic culture, and the renewed interest awaked by the recent celebrations surrounding the 400th anniversary of its publication, ensure a continued popular interest in the *Quixote*. The illustrations of the adventures continue to captivate the popular imagination and help to bring the book alive to millions of readers. By making the collection available to the public in digital format, we are able to radically increase the resources available to students of the *Quixote* from all walks of life. Our goal, however, is not simply to provide a large pile of pictures for students to peruse, but to further enrich their understanding of both the *Quixote* and of art history by supplementing the illustrations with commentary about the significance of the images, their relationship to the text, and the artistic achievement of the illustrators and engravers who created them.

To adequately support both the scholarly and popular communities that will access the collection, it is not sufficient to merely provide an online catalogue of the illustrations, regardless of the completeness of the collection or the detail of the metadata. Instead, we have identified three levels of information required to successfully present these materials:

1. **Descriptive metadata:** The first level of information includes factual descriptions of the items it describes. For our collection, descriptive metadata is provided for both individual illustrations and the editions in which those illustrations were published. This information includes the title of the image, its physical size and the size of the page it is printed on, the name of the artist and/or engraver, the style of the image, and the printing technique used.

2. **Scholarly commentary:** The second level goes beyond simple descriptions of the items in the collection to provide a critical assessment of those items, their narrative context, and their hermeneutic and aesthetic significance. Within our collection we are developing biographical commentary about artists and engravers and technical and artistic commentary about each individual image.

3. **Hypertextual connections:** The third level consists of the information needed to develop dense hypertextual structures and the tools to effectively navigate them.

A key challenge in constructing collections of this type is developing appropriate strategies for interconnecting the resources being continuously added to the collection. While most traditional editorial approaches require editors to carefully edit each resource by hand, this is beyond the scope of our current resources and would greatly limit the size of our collection, the speed with which it could be made available, and the degree of interlinkages that could be provided.

Accordingly, we are using two strategies for automatically interlinking digital resources. One approach relies heavily on the metadata associated with each digital image. We use this metadata information to automatically discover relationships between existing and new resources and then to generate navigational structures based on those relationships. A prototype of this system has been employed internally for use in a project centered on the music in Cervantes' works [9] (Figure 2). One key application of this approach allows us to integrate multiple

resources, including illustrations, with the textual and narrative structure of the *Quixote* [2]. To accomplish this we have developed a formal taxonomy of the narrative and thematic elements of the *Quixote*, which is then used in cataloging each image and can be used for the associated commentary. This allows us to codify the connections between the illustrations and commentary without requiring a particular version of the text. The second approach to interlinking resources we have developed involves elucidating the internal structure of the textual resources in the collection and using that structure to automatically generate navigational links and to inform visualizations [1]; an approach that generalizes Crane's [4]. In this context, we have developed a toolkit to implement this approach for a collection of historical documents pertaining to Cervantes and his family.

Figure 2: The display of resources pertaining to Cervantes' works is automatically constructed from fragmentary entries. Structured links are generated by the system to allow access to a range of resources associated with key phrases.

The hypertextual collections created for our purposes create separately rooted structures over a common set of interlinked elements. For example, in the music collection referred to above, natural collection roots include the compositions, the composers of the pieces, the instruments associated with the pieces, and the texts that refer to the pieces. Each rooted collection includes information specific to the collection (e.g., biographies of the composers and musical scores are associated with their respective collections), but ultimately collections are cross-linked (e.g., between compositions and composers, between instruments and compositions). The texts provide a unifying framework that draws the other components together, yet the texts themselves represent a distinct collection. The structures in the textual iconographic collection show similar relationships, with the added complexity implicit in the three-level categorization of collection information.

In conclusion, by adding detailed scholarly commentary in addition to descriptive metadata and to providing sophisticated tools for automatically building a hypertextual structure into the collection, we are able to provide a new resource that better meets the research needs of scholars and to assist the general public in understanding and appreciating the significance both of the *Quixote* itself, as well as its reception and visual history. Our approach allows the readers of scholarly commentaries more direct access to the primary source materials used to develop and support those commentaries [5] [6,11]. Conversely, it allows individuals focusing on the primary materials access to secondary scholarly works to better understand a variety of reading perspectives and to explore and formulate their own interpretations of these unique materials. Moreover, the nature of the collection and its information relationships has required the addressing of canonical hypertextual structuring issues. Taken as a whole, our collection and hypertextual archive opens previously unavailable opportunities for scholarly study of the *Quixote* and its unique literary, cultural and iconic status.

Footnotes

1. http://cervantes.tamu.edu/
2. In 2001, the Texas A&M University Cushing Memorial Library began an ongoing effort to assemble a rare book collection comprised of all significant illustrated editions of the *Quixote*. Currently this collection contains 413 editions in 15 languages: Don *Quixote* Illustrated: An Exhibit in Celebration of the 4th Centenary of the *Quixote*, 1605-2005. Texas A&M University; Austin: Wind River Press, 2005.
3. John Harthan points out that "book illustration is like a hand-mirror in which one can see reflected great historical events, social changes and the movement of ideas down the centuries" (Harthan, 1981).
4. Indeed, Harthan states that "a history of modern book illustration could almost be written in terms of this perennially popular classic alone" (Harthan, 1981).
5. This echoes our previous work in developing an electronic variorum hyperedition that gives readers access to the primary textual sources used to reach editorial conclusions, thus allowing readers to form their own opinions about which of several variants more likely constitutes the correct text of a disputed passage. http://www.csdl.tamu.edu/cervantes/V2/variorum/index.htm

References

1. **Audenaert, Michael Neal.** 2005. *"Feature Identification Framework and Applications."* M.S. Thesis, Texas A&M University.
2. **Audenaert, N., Furuta, R., Urbina, E., Deng, J., Monroy, C., Sáenz, R., and Careaga, D.** (2005) Integrating Collections at the Cervantes Project. *Proceedings of the Joint Conference on Digital Libraries,* JCDL'05, Denver. 287-288.
3. **Audenauert, N., Furuta, R., Urbina, E., Deng, J., Monroy, C., Sáenz, R., and Careaga, D.** (2005) Integrating Diverse Research on a Digital Library Focused on a Single Author (Cervantes). *Lecture Notes in Computer Science: Research and Advanced Technology for Digital Libraries: 9th ECDL,* Vienna. 151-161.
4. **Crane, G., Wulfman, C. E., and Smith, D. A.** (2001) "Building a Hypertextual Digital Library in the Humanities: a Case Study on London. *Joint Conference on Digital Libraries*, JCDL01, Roanoke. 426-434.

5. **Cervantes, Miguel de. Don** *Quijote de la Mancha*. English translation by Edith Grossman (New York: HarperCollins, 2003).

6. **Kochumman, R., Monroy, C., Furuta, R., Goenka, A., Urbina, E., and Melgoza, E.** (2001) Towards an Electronic Variorum Edition of Don Quixote. *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*. 444-445.

7. **Harthan, J.** (1981) *The History of the Illustrated Book: the Western Tradition*. New York: Thames and Hudson.

8. **McGann, J.** (1997) The Rationale of Hypertext. In Sutherland, K. (ed.): *Electronic Text: Investigations in Method and Theory.* Oxford UP, New York. 19-46.

9. **Pastor, J. J.** *Música y literatura: la senda retórica. Hacia una nueva consideración de la música en Cervantes*.(2005) Doctoral Dissertation, Universidad de Castilla-La Mancha.

10. **Urbina, E.,** et al. (2004) "Iconografía textual del *Quijote*: repaso y nueva aproximación de cara al IV centenario." *Le mappe nascoste di Cervantes. Actas Coloquio Internacional de la Associazione Cervantina di Venecia* (Venice, 2003). Carlos Romero, ed. Treviso: Edizioni Santi Quaranta. 103-114.

11. **Urbina, E.**, General Editor and Director. Electronic variorum edition of the Quixote. Cervantes Project, Texas A&M University. http://cervantes.tamu.edu/V2/variorum/index.htm (accessed November 15, 2005).

12. **Urbina, E., Furuta, R., and Smith, S. E.** (2005) Visual Knowledge: Textual Iconography of the *Quixote,* a Hypertextual Archive. *Proceedings Abstracts ACH/ALLC 2005,* Association for Computers in the Humanities. University of Victoria (Canada). http://web.uvic.ca/hrd/achallc2005/abstracts.htm (accessed November 15, 2005).

13. **Urbina, E.** (febrero-marzo 2005) Don *Quijote*, libro ilustrado. *Contrastes* (Valencia) 38, Special issue: *"Quijote* intesdisciplinar*"*. 37-41.

14. **Urbina, E.** (2005) Iconografía textual e historia visual del *Quijote*. *Cervantes y el pensamiento moderno*. José Luis González Quirós and José María Paz Gago, eds. Madrid: Sociedad Estatal de Conmemoraciones Culturales (in press).

15. **Urbina, E.** (2005) Visual Knowledge: Textual Iconography of the *Quixote*. *Don Quixote Illustrated: Textual Images and Visual Readings*. Eduardo Urbina and Jesús G. Maestro, eds. (Biblioteca Cervantes 2). Pontevedra: Mirabel Editorial. 15-38.

Le livre « *Sous la loupe* ». Nouvelles métaphores pour nouvelles formes de textualité électronique

Florentina VASILESCU ARMASELU

Littérature comparée, Université de Montréal, Canada

1. Introduction

Les expériences sur support électronique de Joyce (1989), Moulthrop (1995), Jackson (1995), Amerika (1997) ont attiré l'attention sur le potentiel du médium électronique comme médium de création littéraire. Elles ont également marqué l'avènement d'un nouveau type d'esthétique, s'intéressant principalement au remplacement éventuel du livre imprimé par le livre électronique (Bolter, 1991; Landow, 1997; Birkerts, 1994), au dilemme de la fin du livre et du livre sans fin (Douglas, 2000), à la capacité prospective de l'hypertexte (Joyce, 2000) ou de la littérature ergodique (Aarseth, 1997) et à la réconciliation de l'immersion et de l'interactivité comme idéal artistique de la «littérature basée sur le langage» (Ryan, 2001).

A la différence des interfaces antérieures (tablettes, rouleaux, codex), l'interface électronique suppose l'existence de deux types de textes. D'un côté, le *texte* destiné au lecteur ou écrit par l'auteur, d'un autre côté, le *code* écrit dans un langage de programmation ou de balisage et déterminant le fonctionnement de l'interface. Cette interaction entre le *texte visible* et le *texte caché,* entre les possibilités d'expression pratiquement illimitées du langage naturel et les contraintes du langage de programmation, représenterait un des éléments centraux dans la production de nouvelles formes textuelles électroniques. Il s'agit alors d'une dépendance à double sens : les codes existants aident à la création de nouvelles formes textuelles et ces formes inspirent à leur tour le développement de nouveaux types d'encodage de plus en plus souples.

2. Enjeux théoriques et pratiques

D'une perspective théorique, notre projet porte principalement sur l'interaction littérature - science - technologie numérique, écriture - programmation, visuel - textuel, dans la création de nouvelles formes textuelles sur support électronique. Nous nous intéressons aussi aux implications de cette interaction sur le domaine de la critique littéraire et notamment aux enjeux d'une mise en relation de l'esthétique du médium électronique avec d'autres champs de la critique contemporaine (*close reading, new historicism,* critique génétique) et des sciences cognitives ou de l'information.

Du point de vue pratique, le projet consiste à expérimenter une nouvelle forme de navigation et d'expression, exploitable par l'intermédiaire de l'ordinateur. Plus précisément, il s'agit de la conception d'un nouveau type d'interface (un éditeur construit dans le langage de programmation Java et utilisant des textes annotés en XML), permettant à l'utilisateur d'augmenter ou de diminuer le degré de détail du texte par une procédure qui rappelle « l'effet de loupe » et la géométrie des fractales. La principale caractéristique de ce type de textualité serait la structuration par niveaux de profondeur, accessibles par des opérations de *zoom-in* et *zoom-out,* le texte de chaque niveau se retrouvant augmenté sur les niveaux subséquents.

Inspirée par la construction fictionnelle de Stephenson (2000) et par le concept de *fractale* de Mandelbrot (1983), notre démarche se propose d'explorer les nouveaux types de rapports auteur - texte - médium - lecteur, de personnages, d'intrigue, de stratégie narrative, de structuration des connaissances ou d'organisation informationnelle suscités par ce genre de développement conceptuel et textuel à plusieurs «échelles». De ce point de vue, notre intérêt porte sur les applications possibles de cette forme d'écriture et de lecture «sous la loupe», impliquant un parcours à mi-chemin entre la *linéarité* et la *non-linéarité*, entre l'*immersion* et l'*interactivité*, caractéristiques habituellement reliées par les théoriciens du médium électronique au *livre imprimé* et respectivement à l'*hypertexte*. Dans ce qui suit, nous reprenons brièvement quelques exemples.

3. Applications possibles

Une première application aurait comme point de départ le concept de *close reading,* une méthode

d'analyse qui examine, à partir d'extraits de texte, le style, les nuances de tons, les stratégies rhétoriques, les hypothèses de nature philosophique ou socio-linguistique (Gallagher et Greenblatt, 2000), et le concept de *new historicism,* une approche qui met en lumière certains aspects d'une œuvre, en essayant de reconstruire son contexte culturel et historique (Greenblatt, 2004). Ce que l'éditeur « sous la loupe » permettrait dans ce cas, est une sorte de *fusion* du texte avec l'analyse historique ou littéraire, développée sur plusieurs niveaux de détails. Par un procédé d'expansion à différentes échelles, on pourrait passer ainsi graduellement du texte littéraire au contexte historique et culturel qui l'a engendré ou imaginer l'outil comme un instrument d'annotation, faisant fusionner le texte, les commentaires critiques et les notes marginales, dans un processus complexe qui rejoint la lecture, la relecture et l'écriture.

Une deuxième application concernerait le domaine de la critique génétique qui s'intéresse aux mécanismes de production textuelle, aux secrets de l'atelier, du laboratoire d'écriture (Grésillon, 1994) et aux rapports existant entre l'écriture et son contexte culturel, entre l'acte de lecture et l'écriture en «train de se faire» (D'Iorio et Ferrer, 2001). De ce point de vue, une analyse «sous la loupe» impliquerait une organisation par niveaux, reliant la forme définitive, les différentes variantes manuscrites, jusqu'aux premiers plans esquissant l'idée de départ du texte. Ce mode d'organisation textuelle mettrait en évidence, de manière graduelle, les hésitations et la dynamique de l'écriture ou les éventuelles traces des lectures antérieures au texte.

Un troisième type d'application viserait l'expérimentation de différentes formes d'expression littéraire. On pourrait imaginer, par exemple, un texte *auto-reflexif*, qui décrive, par une accumulation graduelle de détails, le processus de sa propre création, en retraçant à rebours le trajet de la lecture à l'écriture, de l'expérience vécue à l'expression verbale, et où l'immersion par degrés de profondeur fasse partie de la stratégie narrative. D'autres approches pourraient concerner le développement des personnages dans un texte littéraire. Greenblatt (2004) suggère que, dans *Venus and Adonis*, Shakespeare utilise une technique de rapprochement et d'éloignement du lecteur par rapport aux personnages, en augmentant ou en diminuant le degré de «proximité physique ou émotionnelle». D'un autre côté, Alan Palmer (2003) fait référence au terme de «narration behavioriste» définie comme une narration objective, focalisée sur le comportement des personnages, donc plutôt sur les actions que sur les pensés et sentiments. Pourrait-on imaginer par conséquent une forme narrative permettant une alternance de proximité et de distance par rapport aux protagonistes ou qui, à partir d'une approche béhavioriste, sonde graduellement, les profondeurs psychologiques des personnages ?

D'autres applications du modèle pourraient viser : le domaine pédagogique (structuration des connaissances par degrés de complexité, de l'intuitif vers l'abstrait) ; la construction des dictionnaires ou des encyclopédies (comme collections d'articles « extensibles » adressés à une large catégorie de lecteurs) ; les sciences de l'information (systèmes permettant des réponses par degrés de précision aux interrogations-utilisateur), etc.

4. Conclusion

Bien sûr, il ne s'agit pas de dresser une liste exhaustive des applications du modèle, mais de signaler quelques hypothèses d'étude. Le but de notre présentation serait ainsi une reconsidération de « l'effet de loupe » et de la géométrie des fractales en tant que métaphores pour une nouvelle forme de textualité électronique mettant en lumière certains aspects de la dialectique *texte* / *analyse textuelle ou historique*, *expression écrite* / *idée ou expérience vécue, essentiel* / *détail, intuitif* / *abstrait, précis* / *vague*, dans la production et l'interprétation du texte ou de l'information en général.

The Book "*under the Magnifying Glass*". New Metaphors for New Forms of Electronic Textuality

1. Introduction

The recent experiments on digital support of Joyce (1989), Moulthrop (1995), Jackson (1995) or Amerika (1997) have drawn attention to the potential of the electronic medium as a medium for literary creation. The emergence of the electronic literature has also

determined the coming out of a new form of aesthetics mainly concerning the possible replacement of the printed book by its electronic counterpart (Bolter, 1991; Landow, 1997; Birkerts, 1994), the dilemma of the end of book versus the book without end (Douglas, 2000), the prospective capacity of the hypertext (Joyce, 2000) or of the ergodic literature (Aarseth, 1997), and the reconciliation of immersion and interactivity as a "model for purely language-based literature" (Ryan, 2001).

Unlike the previous interfaces (tablets, roll, codex), the electronic interface supposes the existence of two types of texts: on the one hand, the *text* intended to be read and on the other, the *code* written in a programming or markup language, and determining the performances of the interface. This interaction between the *visible* and the *hidden* text, between the huge potential of expression of the natural language and the constraints of the programming language, represents one of the central elements in the production of new forms of electronic textuality. This kind of dependence implies a double sense relationship: the code supports the creation of new textual forms and these forms can inspire new types of encoding, increasingly flexible.

2. Theoretical and practical assumptions

From a theoretical point of view our project deals with the relationship literature – science – digital technology, writing - programming, visual – textual, in the creation of new forms of electronic textuality. We are also interested in the implications of this kind of interaction upon the domain of literary criticism and especially in the possible relationships between the aesthetics of the electronic medium and other fields of the contemporary criticism (*close reading, new historicism,* genetic criticism), or of the information and cognitive sciences.

The present study consists in an experimental approach dealing with the unexplored possibilities of the electronic support as a medium for textual investigation. Its goal is the construction of a new type of interface (an editor written in Java programming language and using XML annotated texts) allowing the writer and the reader to increase or decrease the degree of detail of the text, by a procedure evoking the "magnifying glass effect" and the fractals geometry. Inspired by the fictional construction of Stephenson (2000) and by the fractal theory of Mandelbrot (1983), this new kind of textuality would be a layout on levels of "depth", accessible by operations of *zoom-in* and *zoom-out*, the text of the most abridged level being reproduced and appropriately augmented on the subsequent, deeper levels. The study deals with the new types of relationship author – text – medium – reader, characters, plot, narrative strategy or knowledge organization determined by this form of "scalable" textual and conceptual structure. The main question addressed by the study would be therefore related to the possible applications of this form of electronic text intended to be written and explored "under the magnifying glass", and implying a halfway between *linearity* and *non-linearity*, *immersion* and *interactivity*, features usually associated with the *printed book* and respectively with the *hypertext*.

3. Possible Applications

The first application would be related to the concepts of *close reading,* a method using short excerpts from a text and carefully examining its style, rhetorical strategies, philosophical and sociological assumptions (Gallagher and Greenblatt, 2000), and of *new historicism,* an approach bringing to light some aspects of a literary work by trying to reconstruct its historical and cultural context (Greenblatt, 2004). The use of the magnifying glass editor would therefore allow a sort of *fusion* of the text with the literary or historical analysis, developed on several levels of detail. We could thus imagine the editor as a tool relating, at different "scales", the literary text with the historical and cultural context having produced it, or with the critical commentaries and marginal notes, in a complex process of reading, re-reading and writing.

The second type of application would concern the field of genetic criticism interested in the dynamics of the process of writing (Grésillon, 1994), and in the intertextual dimension and reading/writing dialectics of the "work in progress" (D'Iorio and Ferrer, 2001). From this point of view, an analysis "under the magnifying glass" would imply a layout on levels relating the definitive form, through different variants, to the first plan sketching the idea of the text. This layout could therefore facilitate the understanding of the gradual dynamics of the act of writing or the recognition of the eventual traces of previous readings. Another type of application may concern the narrative strategies. We can imagine, for instance, a sort of *auto-reflexive* text, conceived as a set of reflections on the act of writing and trying to retrace by details accumulation the path backwards from writing

to reading, from verbal expression to life experience, and involving different "degrees of immersion" as part of the storytelling.

Other approaches could be related to the development of the characters in a literary text. Greenblatt (2004) suggests that in *Venus and Adonis,* Shakespeare uses a technique of approaching or distancing the reader from the characters, by increasing or decreasing his "physical and emotional proximity". On the other hand, Alan Palmer (2003) discusses the term of "behaviorist narrative" defined as an objective narrative focalized on the characters' behavior, i.e. on their actions rather than on their feelings and thoughts. Could we therefore imagine a narrative allowing an alternation of proximity and distance or starting with a behaviorist approach and gradually investigating the psychological depths of the characters?

Other applications of the model could include: the cognitive and pedagogic domain (knowledge organization on levels of complexity, from intuitive to abstract descriptions); the construction of dictionaries and encyclopedias (as collections of expanding articles intended to larger categories of readers); the domain of information science (systems providing several degrees of precision in response to users' queries), etc.

4. Conclusion

Of course, there is not an exhaustive list of possible applications, but rather some directions of study. Our presentation will concern the reconsideration of the "magnifying glass" and fractal geometry as metaphors for a new form of electronic textuality, drawing attention to some aspects of the dialectics: *text / textual or historical analysis*, *written expression / idea or life experience, essential / detail, intuitive / abstract, precision / vagueness,* in textual production and interpretation.

References

Aarseth, E.J. (1997). *Cybertext. Perspectives on Ergodic Literature*. Baltimore, London: The John Hopkins University Press.

Amerika, M. (1997). *Grammatron*. http://www.grammatron.com.

Birkerts, S. (1994). *The Gutenberg Elegies. The Fate of Reading in an Electronic Age*. London, Boston: Faber and Faber.

Bolter, J.D. (1991). *Writing Space. The Computer, Hypertext, and the History of Writing*. New Jersey : Lawrence Erlbaum Associates, Publishers Hillsdale.

D'Iorio P., Ferrer D. (2001). *Bibliothèques d'écrivains*. Paris: CNRS Editions.

Douglas, J.Y. (2000). *The End of Books – Or Books without End? Reading Interactive Narratives*. Michigan: Ann Arbor, The Univerity of Michigan Press.

Gallagher C., Greenblatt S. (2000). *Practicing New Historicism*. Chicago and London: University of Chicago Press.

Greenblatt, S. (2004). *Will in the World. How Shakespeare became Shakespeare*. New York – London: W.W. Norton & Company.

Grésillon, A. (1994). *Eléments de critique génétique. Lire les manuscrits modernes*. Paris: Presses Universitaires de France.

Jackson, S. (1995). *Patchwork Girl by Mary/Shelley and herself*. Watertown, MA: Eastgate Systems,

http://www.eastgate.com/catalog/PatchworkGirl.html.

Joyce, M. (1989). *Afternoon, a story*. Riverrun, Eastgate Systems, http://www.eastgate.com/catalog/Afternoon.html.

Joyce, M. (2000). *Othermindedness. The Emergence of Network Culture*. Michigan: Ann Arbor, The University of Michigan Press.

Landow, G.P. (1997). *Hypertext 2.0. The Convergence of Contemporary Chritical Theory and Technology*. Baltimore, London: The John Hopkins University Press.

Mandelbrot, B. (1983). *The Fractal Geometry of Nature*. New York: W.H. Freeman and Company.

Moulthrop, S. (1995). *Victory Garden*. Eastgate Systems, http://www.eastgate.com/VG/VGStart.html.

Palmer, A. (2003). *The Mind Beyond de Skin*. In *Narrative*

Theory and the Cognitive Sciences. Edited by David Herman. Stanford, California: CSLI Publications.

Ryan, M.L. (2001). *Narrative as Virtual Reality. Immersion and Interactivity in Literature and Electronic Media*. Baltimore and London: The John Hopkins University Press.

Stephenson, N. (2000). *The Diamond Age or A Young Lady's Illustrated Primer*. New York: A Bantum Spectra Book.

If You Build It Will They Come? The Lairah Study: Quantifying the Use Of Online Resources in the Arts and Humanities Through Statistical Analysis of User Log Data

Claire WARWICK
Melissa TERRAS
Paul HUNTINGTON
Nikoleta PAPPA

School of Library, Archive and Information Studies, University College London

The LAIRAH (Log Analysis of Internet Resources in the Arts and Humanities) project aims to determine whether, how and why digital resources in the humanities are used, and what factors might make them usable and sustainable in future. Based at UCL and funded by the UK Arts and Humanities Research Council (AHRC) ICT Strategy scheme, LAIRAH is a year long study which will analyse patterns of usage of online resources through real time server log analysis: the data internet servers collect automatically about individual users. This paper will discuss the findings of our research to date, and the techniques of log analysis as applied to major digital humanities projects. In doing so, we address concerns about the use, maintenance and future viability of digital humanities projects, and aim to identify the means to which an online project may become successful.

Context

Hundreds of projects have been funded to produce digital resources for the humanities. In the UK alone, over 300 of them have been funded by the AHRC since 1998. While some are well known, others have been relatively quickly forgotten, indicating financial and intellectual wastage. Little is known of user centred factors determining usage. (Warwick, 1999) The aim of the LAIRAH study (http://www.ucl.ac.uk/slais/

LAIRAH/) is to discover what influences the long-term sustainability and use of digital resources in the humanities through the analysis and evaluation of real-time use. We are utilising deep log analysis techniques to provide comprehensive, qualitative, and robust indicators of digital resource effectiveness. The results of this research should increase understanding of usage patterns of digital humanities resources; aid in the selection of projects for future funding, and enable us to develop evaluator measures for new projects.

Technical problems that can lead to non-use of digital projects are relatively well understood.(Ross and Gow, 1999) However, evidence of actual use of projects is anecdotal; no systematic survey has been undertaken, and the characteristics of a project that might predispose it for sustained use have never been studied. For example, does the presence in an academic department of the resource creator, or enthusiast who promotes the use of digital resources, ensure continued use? Do projects in certain subject areas tend to be especially widely used? Are certain types of material, for example text or images, more popular? Is a project more likely to be used if it consulted with the user community during its design phase? An understanding of usage patterns through log data may also improve use and visibility of projects.

Project aims and objectives

This project is a collaboration between two research centres at UCL SLAIS: CIBER (The Centre for Information Behaviour and the Evaluation of Research), (http://www.ucl.ac.uk/ciber) and the newly created CIRCAh (The Cultural Informatics Research Centre for the Arts and Humanities) (http://www.ucl.ac.uk/slais/circah/). CIBER members are leading researchers in the use of deep log analysis techniques for the evaluation of online resources. (Huntington et al. 2003) Their strong record in user site evaluation in the health, media and scholarly publishing sectors is now being applied to the arts and humanities. We believe that no one has undertaken log analysis work in this sector: the LAIRAH project therefore offers great opportunities for knowledge and technology transfer.

The LAIRAH project is analysing raw server transactions of online digital resources (which automatically record web site use) and will relate these to demographic user data to provide a comprehensive and robust picture of resource effectiveness. CIBER have developed robust, key metrics and concepts such as site penetration (number of views made in a session), returnees (site loyalty), digital visibility (impact of positioning on usage), search success (search term use and the number of searches conducted) and micro-mining (mapping individual tracks through websites) in order to understand usage in the digital environment and relate this to outcomes and impacts. (Nicholas et al., 2005) By applying this knowledge in the LAIRAH project, we are bringing quantitative and robust analysis techniques to digital resources in the Arts and Humanities.

Methods

Phase 1: Log analysis

The first phase of the project is the deep log analysis. Transaction and search log files have been provided by the three online archives supported by AHRC: the **AHDS** Arts and Humanities Collection (http://www.ahds.ac.uk/); **Humbul** Humanities Hub (http://www.humbul.ac.uk/) and the **Artifact** database for the creative and performing arts (http://www.artifact.ac.uk/), (Humbul and Artifact merged at the end of 2005). This provides rich data for comparing metrics between subject and resource type. The search logs show patterns of which resources users are interested in, and in the case of the AHDS (which provides links through to resources themselves), which ones users go on to actually visit. This project is of limited duration, thus we are not able to analyse logs from individual projects at this stage of funding, given the difficulty of accessing log data from projects which may have limited technical support.

We are analysing a minimum of a year's worth of transaction log data (a record of web page use automatically collected by servers) from each resource. This data gives a relatively accurate picture of actual usage, is seamless, and is easily available, providing: user information on the words searched on (search logs), the pages viewed (user logs), the web site that the user has come from (referrer logs), and basic, but anonymous, user identification tags, time, and date stamps. (Huntington et al., 2002) We have also designed short online questionnaires, covering user characteristics and perceived outcomes, which will be matched to actual search and usage patterns. We are also sharing the results of these and the log data analysis with

another project, also funded under the AHRC ICT scheme, (http://www.ahrbict.rdg.ac.uk/) the RePAH project (User Requirements analysis for Portals in the Arts and Humanities http://repah.dmu.ac.uk/), based at De Montfort and Sheffield universities. This project is studying the use of Portal sites in the arts and humanities, and is testing new prototype portal designs, including applications such as personalisation functions used on commercial portals, to determine whether they are appropriate for humanities users.

As part of the initial phase of the project, we have also carried out a study to determine how humanities users find digital resources and portal sites, when beginning their search from their university library or faculty web pages. This study is described in a separate poster proposal.

Phase 2: Case Studies

Through the above log analysis, we have identified ten projects that have high and low patterns of use across different subject areas and types of content and these are being studied in depth. Project leaders and researchers are being interviewed about project development, aims, objectives, and their knowledge of subsequent usage. Each project is analysed according to its content, structure, and design and whether it has undertaken any outreach or publicity. We are seeking to discover whether projects have undertaken user surveys, and if so how they responded to them and whether they undertook any collaboration with similar projects. We are also asking about technical advice that the project received, whether from institutional support people, from Humanities Computing Centres or from central bodies like the AHDS.

All these measures are intended to determine whether there are any characteristics which projects which continue to be used may share. For example does good technical advice predispose a project to be usable or might contact with potential users prove as important? We shall also interview a small sample of users of each resource about their opinions about the reasons why it is useful for their work. This aspect of the project will be collaborative with another CIRCAh project, which is studying the reaction of humanities users to digital projects: the UCIS project.

We nevertheless recognise that it is also important to study projects that are neglected or underused. We are therefore running a workshop with the AHRC ICT Methods Network to study the possibility of the reuse of neglected resources. A small group of humanities users will be given an opportunity for hands on investigation of a small number of resources and there will be time for discussion of factors that might encourage or deter their future use. We will seek to find out whether their lack of use is simply because users had not heard of a given resource, are whether there are more fundamental problems of design or content that would make the resource unsuitable for academic work.

Findings

Collecting the log data has proved to be an unexpectedly difficult process. This in itself is noteworthy, since it indicates that levels of technical support even for large, government supported portals could still be increased. The situation for individual projects is likely to be even more problematic, and suggests that the issue of long term maintenance and support is one that institutions and funding councils must take more seriously.

We are currently beginning to analyse data, and, by July, we will be in the final quarter of the project. We will therefore be able to report on the results of both the qualitative and quantitative aspects of the study. These should prove valuable to anyone at the conference who is currently running or planning to run a future digital resource for the humanities. Like all other matters in the humanities, building a digital resource which is successful in term of attracting and keeping users is not an exact science. We do not mean to limit the creativity of culture developers by suggesting the application of a rigid list of features to which all future projects must conform. Nevertheless, since resource creators spend such large amounts of precious time, effort and money on making their project a reality, they must surely be keen to see it used rather than forgotten. We aim to suggest factors which may predispose a resource to continued success, in terms of users: a topic of interest to project designers, project funders and users alike.

Acknowledgement

The LAIRAH project is funded by the AHRC ICT Strategy scheme.

References

Huntington P, Nicholas D, Williams, P. (2003) 'Characterising and profiling health web users and site types: going beyond hits'. *Aslib Proceedings* 55 (5/6): 277-289

Huntington P, Nicholas D, Williams P, Gunter B. (2002) 'Characterising the health information consumer: an examination of the health information sources used by digital television users'. *Libri* 52 (1): 16-27

Nicholas D, Huntington P, and Watkinson A. (2005) 'Scholarly journal usage: the results of a deep log analysis.' *Journal of Documentation* 61 (2): 248 – 280.

Ross, S., and Gow, A. (1999) Digital Archaeology, Rescuing neglected and damaged data resources. HATII, University of Glasgow. Online http://www.hatii.arts.gla.ac.uk/research/BrLibrary/rosgowrt.pdf

Warwick, C. (1999) 'English Literature, electronic text and computer analysis: an unlikely combination?' Paper presented at *The Association for Computers and the Humanities- Association for Literary and Linguistic Computing, Conference*, University of Virginia, June 9-13.

Code, Comments and Consistency, a Case Study of the Problems of Reuse of Encoded Texts

Claire WARWICK
School of Library, Archive and Information Studies, University College London

George BUCHANAN
Department of Computer Science, University of Swansea

Jeremy GOW
Ann BLANDFORD
UCL Interaction Centre, University College London

Jon RIMMER
School of Library, Archive and Information Studies, University College London

Introduction

It has long been an article of faith in computing that when a resource, a program or code is being created, it ought to be documented. (Raskin, 2005) It is also an article of faith in humanities computing that the markup should be non-platform-specific (e.g. SGML or XML). One important reason for both practices is to make reuse of resources easier, especially when the user may have no knowledge of or access to the original resource creator. (Morrison et al, nd, chapter 4)

However, our paper describes the problems that may emerge when such good practice is not followed. Through a case study of our experience on the UCIS project, we demonstrate why documentation, commenting code and the accurate use of SGML and XML markup are vital if there is to be realistic hope of reusing digital resources.

Background to the Project

The UCIS project (www.uclic.ucl.ac.uk/annb/DLU-sability/UCIS) is studying the way that humanities researchers interact with digital library environments.

We aim to find out how the contents and interface of such collections affect the way that humanities scholars use them, and what factors inhibit their use. (Warwick, et al., 2005) An early work-package of the project was to build a digital text collection for humanities users, delivered via the Greenstone digital library system. We chose to use texts from the Oxford Text Archive, (OTA) because this substantial collection is freely available and contains at least basic levels of XML markup. However this task was to prove unexpectedly difficult, for reasons that extend beyond the particular concerns of UCIS.

Findings

On examination of a sample of the files, we found that although they appeared to be in well formed XML, there were many inconsistencies in the markup.

These inconsistencies often arise from the electronic history of the documents. The markup of older (Early and Middle) English texts is complex, and many of the problems stem from succeeding revisions to the underlying content. One common early standard was Cocoa markup, and many of the documents still contain Cocoa tags which meant that the files would not parse as XML.

In Cocoa, the (human) encoder can provide tags that indicate parts of the original document, their form and clarity. These tags were retained in their original Cocoa form which was mistaken for potential TEI tags by the processing software. Many characters found in earlier English were encoded using idiosyncratic forms where modern (Unicode or SGML Entity) alternatives now exist. The earlier, Cocoa, form may render the modern electronic encoding unparsable in either XML or SGML.

Another problem with Cocoa markup is that it was never fully standardised, and tags are often created or used idiosyncratically. (Lancashire, 1996) This complicates a number of potential technical solutions (e.g. the use of XML namespaces). Some content included unique tags such as "<Cynniges>": not part of any acknowledged hybrid of the original standard. The nature of this is unclear. It may be an original part of the text, (words actually surrounded by '<' and '>'), a Cocoa tag, or a TEI/SGML/XML tag. The distinction of forms known to a modern TEI/XML document is straightforward; the distinction between Cocoa and SGML/XML is not possible in this context.

Even parts of the same document used the same tag inconsistently. For example, distances (e.g. "ten lines of space") may be rendered in numeric form ('10 lines') or textual form 'ten lines') and distance units may be given in full or abbreviated.

One common character notation was '&&' to represent the 'Thorn' character (in upper case) and '&' to represent the same character in lower case. This was interpreted as a SGML/XML entity, but parsers were unable to successfully interpret the original scheme. Furthermore, as the SGML/XML format was used in other parts of the document, even a bespoke parser could not successfully disambiguate the intention of every occurrence of the '&' character. Thus, content is effectively lost. Other characters remain in forms such as '%' for '&' or 'and' –because of the original special use of '&'. Such characters thus remain unintelligible to an SGML or XML document reader. Given these complications, it is often impossible for a computer to determine the proper form of the document without human intervention, making automatic processing and indexation impossible.

As Giordano (1995) argues, 'No text encoded for electronic interpretation is identifiable or usable unless it is accompanied by documentation'. Yet in none of these cases did we find that markup decisions had been documented, nor was the code commented. The OTA supplied each file with a TEI header, which provides some basic metadata about its creation. However, the header was intended to act as the kind of metadata that aids in resource discovery, rather as code books were used to find a specific social science dataset on a magnetic tape. The <encodingdesc> element is not mandatory, and was intended to explicate transcription practices rather than detailed markup decisions. (Giordano, 1995) We certainly did not find any examples of attempts to elucidate markup schemes in the headers. Documentation was also not available for any of the files we looked at. Though the OTA strongly encourage depositors to document their work, they do not mention markup specifications as an element of basic documentation, so even documented files might not have provided the information we needed. (Popham, 1998). We were therefore forced to attempt to reconstruct decisions made from visual examination of each file.

Despite the help of the OTA with cleaning up the data, the task proved so large that we had to abandon the use of these files. We have therefore used commercially

produced resources, with the permission of Chadwyck Healey limited. The advantage of using their material for our project was that the markup is consistent, has been documented and conforms to written specifications.

Conclusions

It is to be hoped that simply by drawing attention to some of the problems that may occur in reuse, our work will cause resource creators to take seriously the importance of documentation and consistency. We have reported a case study of one UK-based repository, but since the OTA is one of the most reputable sources of good quality electronic text in the world, our findings should be of interest to the creators and users of other electronic texts well beyond this particular example. Not all electronic texts are of such high quality, nor are they always collected by an archive, and so such considerations become even more important when texts are made available by single institutions such as libraries, university departments or even individual scholars.

One of the objectives of the Arts and Humanities Data Service (the organisation of which the OTA is a part) since its foundation has been to encourage the reuse of digital resources in humanities scholarship. Yet our experience has shown that the lack of consistency and documentation has made this task almost impossible. The advantage of markup schemes such as XML should be that data is easily portable and reusable irrespective of the platform within which it is used. Yet the idiosyncratic uses of markup that we found have almost negated this advantage.

The creators of the resources probably thought only of their own needs as researchers and were happy with markup that made sense to them. It is still common for projects that use TEI to create their own extensions, without necessarily documenting them. Unlike computer scientists, whose collaborative research practices make them aware of the importance of adhering to standards and conventions that make their code comprehensible, humanities scholars are rewarded for originality, and tend to work alone. Research paradigms do not oblige scholars to think about how their work might be reused, their data tested, or their resource used to further research collaboration. One recommendation that follows from our work is that humanities scholars should at least take advice from, and ideally collaborate with, computer scientists or technical specialists, whose collaborative research practices make them aware of the importance of adhering to standards and conventions that make their code comprehensible.

This might not matter if the creators of a resource are its only users, but given the intellectual and monetary cost of resource creation, their authors ought at least to be aware of the possible implications of applying idiosyncratic markup without comments or documentation. This paper provides the evidence of just such consequences.

References

Giordano, R. (1995) 'The TEI Header and the Documentation of Electronic Texts.' *Computers and the Humanities*, 29 (1): 75-84.

Lancashire, I. (1996) Bilingual Dictionaries in an English Renaissance Knowledge Base. Section 3. Computers in the Humanities Working Papers. University of Toronto. Available at http://www.chass.utoronto.ca/epc/chwp/lancash1/lan1_3.htm

Morrison, A. Popham, M and Wikander, K. (no date) *Creating and Documenting Electronic Texts: Guide to Good Practice*, AHDS publications. Available at http://ota.ahds.ac.uk/documents/creating/

Popham, M. (1998) *Oxford Text Archive Collections Policy - Version 1.1*, AHDS Publications. Available at http://ota.ahds.ac.uk/publications/ID_AHDS-Publications-Collections-Policy.html

Raskin, J. (2005) '**Comments are more important than code'.** *Queue* 3 (2): 64-66.

Warwick, C., Blandford, A., Buchanan, G. & Rimmer, J. (2005) 'User Centred Interactive Search in the Humanities.' *Proceedings of 5th ACM/IEEE-CS joint conference on Digital libraries*. **New York: ACM press. p. 400.**

The Three-Dimensionalisation of Giotto's 13th-Century Assisi Fresco: *Exorcism of the Demons at Arezzo*.

Theodor G. WYELD

IEP, ITEE,
The University of Queensland, Aust.
twyeld@itee.uq.edu.au

Giotto's Fresco

Giotto's thirteenth-century fresco *Exorcism of the Demons at Arezzo* in the Church of San Francesco in Assisi is often referred to as marking the transition from the flattened medieval Byzantine ritualised image to the more spatially realistic perspectives of the Renaissance proper (Damisch, 1994; Edgerton, 1991; Gombrich, 2000; Kemp, 1990; Kubovy, 1989; Panofsky, 1991) (see figure 1). His achievements were recognised by his contemporaries such as Dante and Cennini, and his teacher Cimabue. He had a profound influence on Florentine painting in general and inspired the generation of artists that followed such as Masaccio and Michelangelo. In this, the tenth panel of a series of twenty eight frescos, we see an awkward (by modern standards) attempt to depict depth on a two-dimensional surface. His frescos attempted to illustrate the natural world with depth cues such as receding lines and chiaroscuro shading techniques. He also broke with the tradition of strictly depicting size relationships between people in a scene according to their hegemonic hierarchy. Instead, Giotto illustrated a spatial hierarchy between objects in a scene – including people. On the left we see the cathedral of San Donato (now the Diocesan Museum) with St Francis and Brother Sylvester attempting to drive out the demons aloft over the city, to the right of the fresco. The cathedral has been constructed with lines receding to the left suggesting distance. This is incongruous, however, with the city buildings to the right which have their diminishing lines marching to the right (see figure 2). Hence, as a complete composition, it does not portray the truly unified perspectival space we are more accustomed to that came later in the Renaissance. Nevertheless, in his clumsy way Giotto had established a sense of depth in his paintings which would have been just as profound to the uninitiated as any photograph we could produce of the scene today.

Figure 1. Giotto's thirteenth-century fresco Exorcism of the Demons at Arezzo in the Church of San Francesco in Assisi (KFKI, 2004).

Was Giotto's depiction of depth really that clumsy though? Perhaps by today's mathematically precise algorithmic computer-generated perspectives it is. Or, perhaps Giotto was not attempting to depict a realistic scene, as much as later Renaissance paintings would, but simply hinting at the spatial arrangements of the city of Arezzo, the cathedral and surrounding countryside? The city of Arezzo depicted in Giotto's fresco dates back to the sixth century BC. At the time, the city was situated on the top of the Donato Hill where we can now find the Prato Gardens and the fifteenth-century Medici Fortress. Between the Cathedral and the Fortress was a vast natural depression. The cavity has since been filled in to construct the Prato Gardens. Many of the original features are present in his fresco.

Figure 2. Spatial analysis of Giotto's fresco.

The Three-Dimensionalisation of Giotto's Fresco

To determine how accurate, or rather what are the spatial interrelationships between the various buildings and landscape depicted in his fresco, the author of this paper embarked upon its three-dimensionalisation. In the spirit of the endeavour that Giotto himself took, I worked directly from an image of the fresco meticulously reconstructing each individual element in it. In an architectural sense, one would normally work from plans and project the three-dimensional volumes over them. In this case, there are no plans, so a different strategy had to be found. Working as Giotto himself may have done, the method adopted was based on one of simple proportions. A geometrical analysis of the image produced a grammar of sorts that appeared to be consistent across most of the city buildings. For example, there are simple geometric relationships between the window openings and the masonry that surrounds them (1:1, 1:2, 1:3, and so on). The next relationship is the relative heights. If we assume most of the main buildings were at most two storeys then, judging by their relative heights, four distinct ground levels emerge that these buildings are perched on. In the thirteenth century the extent of building technology typically only allows for a maximum of four storeys in domestic construction. Hence, the towers are typically twice the height of the average dwelling. From here a topology emerges consistent with the actual site in Arezzo today.

Analysis of the Model

The next stage was to assemble the city buildings. While at first, they just seem to be a dense agglomeration, by their three-dimensional modelling and organisation, and according to the picture analysis, an order emerges. This was revealed as much by the reconstructive modelling as it was by analysis of the picture itself. For example, clearly some buildings were in front of others, and others to the side of these. Hence, once the proportions were determined gaps could be detected between buildings (when viewed from above) that are not immediately obvious in Giotto's original fresco. Indeed, an overall layout for the city (that part that was visible in Giotto's fresco) includes what appears to be open spaces connected by access alleys between buildings. At one stage of the project a significant city square was revealed by the overall layout. However, when the different levels of the buildings were taken into account, this turned out to be simply a steep section of land that could not be built on (see figure 3).

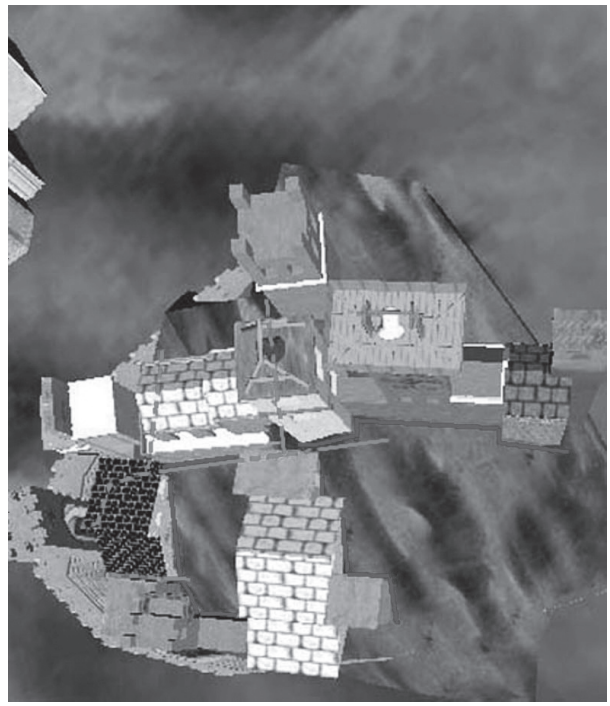


Figure 3. Bird's-eye-view of the three-dimensionalisation of Giotto's fresco. Note the open spaces and alleys between buildings.

Other details were revealed through this procedure too. For example, all the towers house a bell. This, in conjunction with the airborne demons, the gesticulations of the priests, and citizens gathered at the city gates, suggests that Giotto's depiction of so many bell towers had a purpose. The 3D model is an interactive spatial model that can be navigated in real-time. As such, within the multimedia reconceptualisation of his fresco it was possible to add the sounds of the bells. This adds a level of meaning to the fresco that is not apparent from the picture alone (see figure 4).

Figure 4. Screen grab of interactive real-time navigable model of Giotto's fresco three-dimensionalised.

What was Revealed

Together, the details derived from the analysis of Giotto's fresco, the spatial characteristics of its subsequent three-dimensionalisation, and the addition of bell ringing provide another level of experience and understanding of Giotto's work that he could not have anticipated. More than this, it exposes Giotto's spatial reasoning to be more developed than many had thought previously. The fact that such an accurate reconstruction of his fresco can be produced suggests greater insights into spatial relationships between the objects in his fresco were present than previously reported. However, as much of the original city of Arezzo was dismantled in the fifteenth century to build the Medici fortress it is almost impossible to determine if the arrangement of buildings in Giotto's city of Arezzo are a natural recording. Nevertheless, we do know that he was one of the new breed of naturalists actively seeking greater clarity in illustrating the world around him, therefore it is reasonable to assume that his pictorial depiction of the city is much more than a simple, stylised, ritualistic, scenography.

References

Damisch, H. (1994). *The Origin of Perspective***.** Goodman, J. (trans.) MIT Press, Cambridge, Mass, London, England.

Edgerton, S. Y. (1991). *The Heritage of Giotto's Geometry: Art and Science on the Eve of the Scientific Revolution***.** Cornell University Press, Ithaca, London.

Gombrich, E. H. (2000). *Art and Illusion: A Study in the Psychology of Pictorial Representation***.** Princeton University Press, Princeton and Oxford.

Kemp, M. (1990). *The Science of Art: Optical Themes in Western Art from Brunelleschi to Seurat***.** Yale University Press, New Haven and London.

Központi Fizikai Kutató Intézet (KFKI). *Web Gallery of Art***.** source: www.wga.hu/ Dec, 2004.

Kubovy, M. (1989). *The Psychology of Perspective and Renaissance Art***.** Cambridge University Press, Cambridge.

Panofsky, E. (1991). *Perspective as Symbolic Form***.** Wood, C., (Trans.), Zone Books, New York.

Connecting Text Mining and Natural Language Processing in a Humanistic Context

Xin XIANG
John UNSWORTH

Graduate School of Library and Information Science, University of Illinois, Urbana, Champaign1 Introduction

Recent integration of advanced information technology and humanistic research has seen many interesting results that are brand new to traditional humanistic research. In the NORA project this integration was largely exemplified. In an effort to produce software for discovering, visualizing and exploring significant patterns across large collections of full-text humanities resources in digital libraries and collections, NORA project features the powerful D2K data mining toolkit developed by NCSA at University of Illinois, and the creative Tamarind preprocessing package developed by University of Georgia. In NORA project, D2K and Tamarind need to talk to each other, among other relevant components. The idea behind this connection is as follows:

- Make use of existing efforts and try to avoid duplication. Natural language processing, such as part-of-speech tagging, word sense disambiguation, and bilingual dictionary creation, has long been recognized as an important technique for text data mining (Amstrong, 1994). D2K has proved to be an effective and comprehensive data mining toolkit, and Tamarind prepares data gleaned from large-scale full text archives. Getting them work together is an easy and time-saving way of achieving the goal of NORA.
- Separate different tasks according to institution makes the multi-institutional project easier. D2K has been developed and used in several institutions within University of Illinois, and Tamarind was developed in University of Georgia for simplifying primary text analysis tasks. This separation keeps each institution focusing on a relatively independent module that they have the most experience with.
- Prepare information about tokens once and for all. Natural language processing tasks prove time-consuming and computation-intensive. Separating these tasks from data mining part of this project obviates D2K toolkit from performing basic data analysis every time it runs, thus streamlining the whole process.

The problem, however, is that D2K and Tamarind are developed using different programming languages and have different communication mechanisms. To put them together requires reconcilement and restructuring at both sides. This has eventually been achieved in a prototype application, where a collection of Emily Dickinson's poems is classified as either "hot" (erotic) or "not hot" based on the language used (Kirschenbaum, 2006).

As the size of data increases, the problem of scalability emerges. The huge size of many humanistic collections will make unrealistic the solution of storing all the tables in the database. A perfect method to address this problem has not been found, and content presented here demonstrates how we approach the text mining problem in the prototype when the size of collections is not very large.

2 The D2K Toolkit

D2K - Data to Knowledge is a flexible data mining and machine learning system that integrates analytical data mining methods for prediction, discovery, and deviation detection, with information visualization tools (D2K). It provides a graphic-based environment where users with no knowledge in computation and programming can easily bring together software functional modules and make an itinerary, in which a unique data flow and a task are performed. These modules and the entire D2K environment are written in Java for maximum flexibility and portability.

The data mining and machine learning techniques that have been implemented in D2K include association rule, Bayes rule, support vector machine, decision tree, etc. These techniques provide many possibilities of classifying collections available to this project, like hundreds of Emily Dickinson's poems.

Although D2K has the ability of performing basic natural language processing tasks, it is still beneficial to delegate those tasks to a toolkit that is specifically designed to do this, i.e., Tamarind.

3 Gate and Tamarind

Both D2K and Tamarind use Gate as their fundamental natural language processing toolkit. Gate has been in development at the University of Sheffield since 1995 and has been used in a wide variety of research and development projects (Gate).

Tamarind is a text mining preprocessing toolkit built on Gate, analyzing XML-based text collections and putting the results into database tables (Downie 2005). It serves as a bridge between Gate and D2K, and connects them through the use of persistent database. It supports JDBC-based data retrieval, as well as SOAP-based language-independent APIs.

Table 1 shows a typical table in Tamarind database. The "xpath" field contains the location of a token in the TEI document in terms of XPath expression, "doc_id" is the unique ID of the TEI

document, while "t_type_id" is the part-of-speech tag. Based on this table, some statistical characteristics of tokens, like term occurrence (term frequency), co-occurrence and document frequency, could be generated, thereby obviating the data mining toolkit (D2K) from performing the data-preparing task.

xpath	doc_id	pos_id	t_type_id	token_id
/TEI.2[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/title[1]/Token[1]	1	1	1	1
/TEI.2[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/title[1]/Token[2]	1	2	2	2

Table 1: A Tamarind Table

After the whole collections is parsed and analyzed, the information related to the position, part of speech and type of each token is stored in a PostgreSQL database for future access. The Tamarind application exposes these information so that D2K as a client can connect and retrieve them through JDBC (Java Database Connectivity) or SOAP (Simple Object Access Protocol).

4 NORA Architecture

Several issues were raised as to how to effectively and efficiently connect physically and institutionally distributed components in the NORA project. For example, should Tamarind expose its data to client through Java API (as a Java JAR file) or SOAP API (through Web service)? Should Tamarind just provide raw data like that in the previous table or something more advanced and complicated like the frequently used TF-IDF value? Is D2K responsible for converting the database table to a data structure more convenient for D2K to handle, like D2K table? How can the user requests be conveyed to D2K in a user-friendly and compact fashion?

Experiments and discussion eventually led to the adoption of JDBC-based data retrieval and SOAP-based Web service for user request delivery. Although SOAP-based Web service providing more advanced and platform-independent API interface is a good choice for delivering Web-based requests, it seems inefficient to transmit large amount of data, like the occurrences of all tokens in the whole collection, through HTTP protocol, especially when the data store and the text mining application do not reside on the same host. This, however, does not exclude the possibility of implementing some not-so-data-intensive APIs, like metadata retrieval, through SOAP in the future.

Table 2 gives sample data pairs pulled out of Tamarind database. It is a list of which token occurs in which document and is generated by a join of several tables in the Tamarind database.

document	token
Seaf709v1-tbh.xml	with
Seaf709v1-tbh.xml	that
Seaf709v1-tbh.xml	of
Seaf709v1-tbh.xml	joseph
Seaf709v1-tbh.xml	in
Seaf709v1-tbh.xml	egypt

Table2 : Data (document-token pairs) from Tamarind Database

After data about tokens is pulled out of the Tamarind database, it is converted to a structure called "D2K table" which is convenient for the D2K toolkit to handle. Actually the D2K table is the restructuring of the token-document pairs taken from the database as a matrix containing the occurrences of each token in each document. Table 3 gives an example. Depending on the collection, it could contain hundreds of rows and thousands of columns.

	with	that	of	joseph	in
Seaf709v1-tbh.xml	0	0	1	1	0
Seaf709v2-tbh.xml	1	0	1	1	0
Seaf709v3-tbh.xml	0	2	1	1	1
Seaf709v4-tbh.xml	0	0	0	2	0
Seaf709v5-tbh.xml	1	0	0	0	1

Table 3: A D2K Table

This D2K table is ready for evaluation by several common machine learning techniques, like naive Bayes and support vector machine. For the Dickinson prototype, naive Bayes algorithm is used and an overall classification accuracy of over 70% is achieved. In the prototype, the D2K toolkit is launched by Infovis, an information visualization toolkit, through Web service.

Figure 1 depicts the control flow of the whole prototype system.

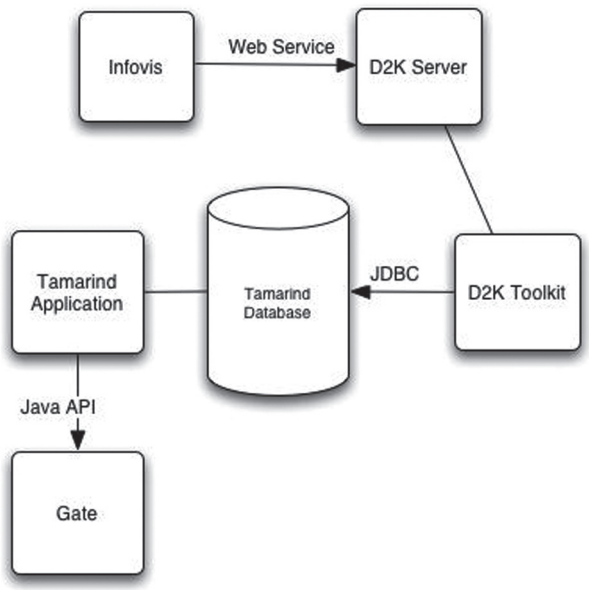


Figure 1. NORA Architecture

References

Armstrong, S. (1994). *Using Large Corpora*. MIT Press.

Kirschenbaum, M. Plaisant, C. Smith, M. Auvil, L. Rose, J. Yu, B. and Clement, T. (2006) *Undiscovered public knowledge: Mining for patterns of erotic language in Emily Dickinson's correspondence with Susan (Gilbert) Dickinson*. ACH/ALLC 2006.

http://alg.ncsa.uiuc.edu/do/tools/d2k.

http://gate.ac.uk.

Downie, S. Unsworth, J. Yu, B. Tcheng, D. Rockwell, G. and Ramsay S. (2005) *A revolutionary approach to humanities computing?: Tools development and the D2K data-mining framework*. ACH/ALLC 2005.

Toward Discovering Potential Data Mining Applications in Literary Criticism

Bei YU
John UNSWORTH
University of Illinois at Urbana - Champaign

1. Introduction

Over the past decade text mining techniques have been used for knowledge discovery in many domains, such as web documents, news articles, biomedical literature, etc. In the literary study domain, some data mining applications have emerged, among which document categorization may be the most successful example (Meunier 2005). But the overall progress of computer assisted literary study is not significant.

The goal of this research is to discover more potential data mining applications for literary study. The basic belief underneath our research is that in order to better adapt data mining techniques to literary text, one has to grasp the unique characteristics of literary research and to leverage its uniqueness and its similarity with data mining. Buckland (Buckland, 1999) claimed that vocabulary is a central concept in information transition between domains. Comparing the vocabularies between the corpora in different domains may shed light on discovering the similarity and difference in the research activities between these domains. So we propose a 3-stage approach to map research activities between data miners and literary scholars as reflected in the vocabulary use in their research publications. Stage 1 is to investigate literary scholars' unique research activities by verb analysis and topic analysis in critical literature, and see if any available data mining techniques can be applied to assist the scholars in these activities. Stage 2 is to investigate the mainstream data mining practices and the representations of the discovered knowledge by keyword analysis in data mining literature, and see if they also appear in critical literature setting. The shared research activities and knowledge representations will suggest some research problems on which data mining experts and literary scholars can start their collaboration. The two stages are complimentary to each other rather than sequential. In the last stage, potential literary text mining problems are summarized into a list of questions, and some literary scholars are interviewed to verify if these applications are useful and which of them can be specified to be ready for text mining.

Up to date we have finished the first two stages. We will be interviewing 5-10 literary scholars between now and the conference. The results of the interviews will be included in our presentation at the conference.

2. Corpus Construction

Three corpora have been constructed for the vocabulary use analysis in stage 1 and 2. The first is the data mining corpus (named "KDD") which consists of 442 ACM SIGKDD conference paper abstracts from 2001 to 2005. The ACM SIGKDD conference has been the premier international conference on data mining. The paper titles and abstracts are extracted from the ACM Digital Portal. We do not use full text because it contains too many technical details that are not relevant to literary research.

The second is the literary criticism corpus (named "MUSE") which consists of 84 ELH Journal articles and 40 ALH articles downloaded from Project Muse, all on the subject of the 18th and 19th century British and American literature. The selection is based on the subject indexes assigned by the publisher. The plain text versions are generated by removing all the tags and quotations from the corresponding HTML versions.

The third is the New York Times subset of American National Corpus (named "ANC-NYTIMES") which consists of thousands of news articles with more than 2 million words. This "everyday English" corpus serves as a contrast group to test if the discovered similarities between the research behaviors in data mining and literary study are significant.

3. Stage 1: discovering literary scholars' unique research activities

This stage consists of three steps. Firstly, the plain text MUSE documents are part-of-speech tagged using GATE. Document frequency (DF) and term frequency (TF) serve as the basic indicators for a term's

popularity in a collection. Arbitrary DF is defined as the number of documents that contain the term. Normalized DF is defined as the percentage of the arbitrary DF in the collection (denote as "DF-pcnt"). Arbitrary TF is defined as the term's total number of occurrences in the whole collection. Normalized TF is defined as the proportion per million words (denote as "TF-ppm"). The verbs are cascade sorted by their DF and TF.

A literary scholar picked out some representative verbs (with both DF and TF between 5 and 10) in critical literature setting: "clarifies", "cleared", "Knowing", "destabilizes", "analyzing", "annotated", "juxtaposed", "evaluated", "recapitulates", "merit", "detail", "portraying", and "stemming".

Secondly, a unique MUSE verb list is generated by comparing the verbs in MUSE and ANC-NYTIMES, also cascade sorted by DF and TF. The top 10 unique verbs are "naturalizing", "narrating", "obviate", "repudiate", "Underlying", "misreading", "desiring", "privileging", "mediating", and "totalizing".

Obviously the two verb lists do not overlap at all. Actually, the representative verbs (except "recapitulates") picked out by the literary scholar turn out to be common in ANC-NYTIMES corpus too. After examining the unique MUSE verb list, two literary scholars were surprised to find many unexpected unique verbs, which means their uniqueness is beyond the scholars' awareness.

Thirdly, simple topic analysis shows that many MUSE essays are trying to build connections between writers, characters, concepts, and social and historic backgrounds. As an evidence, 56 out of 84 ELH essays and 24 out of 40 ALH essays titles contain "and" - one of the parallel structure indicator. But genre is the only topic that can be mapped directly to text mining application - document categorization.

In conclusion, literary scholars are not explicitly aware of what are the unique research activities at the vocabulary-use level. They might be able to summarize their scholarly primitives as Unsworth did in (Unsworth, 2000), but does not help computer scientist to understand the data mining needs in literary criticism.

4. Stage 2: discovering the mainstream data mining activities and the representations of discovered knowledge in KDD and MUSE corpora

This stage of analysis consists of two steps: 1) extracting keywords from KDD paper titles, identifying mainstream data mining activities and knowledge representations in data mining; and 2) comparing the DFs and TFs of the KDD keywords between KDD, MUSE, and ANC-NYTIMES corpora, identifying the keywords common in both KDD and MUSE but not in ANC-NYTIMES.

In the first step, non-stop words are extracted and stemmed (using Porter Stemmer) from paper titles and sorted only by their TF. 18 out of 102 non-stop stemmed title words with TF>5 are identified as the representative data mining keywords. The left out terms include general terms (e.g. "approach"), technical terms (e.g. "bayesian"), terms about specific data (e.g. "gene"), and terms with different meaning in MUSE (e.g. "tree").

Table 1 compares the frequencies of the 18 words between MUSE and ANC-NYTIMES. It shows that 11 data mining keywords are common in literary essays but not in news articles. Figure 1 visualizes their significant differences in TF-ppm. The 11 keywords stand for models, frameworks, patterns, sequences, associations, hierarchies, classifications, relations, correlations, similarities, and spatial relations. It's not surprising that none of these keywords can be found in MUSE essay titles. The context of the keywords extracted from KDD abstracts and MUSE full text also has little in common.

In the left 7 KDD keywords, "rule", "serial/seri" and "decis" are common in both corpora, "cluster" and "stream" are common in neither of them. Interestingly "network" and "graph(ic)" are much more common in ANC-NYTIMES. It seems literary scholars do not think much in graphic models.

In conclusion, literary scholars are actually "data miners", except that they look for different kinds of knowledge. For example, in terms of pattern discovery, literary scholars look for "narrative patterns", "marriage patterns", "patterns of plot", etc. But data miners concern pattern in a more abstract manner - "sequential patterns", "association patterns", "topological patterns", etc.

5. Stage 3: interview the literary scholars

to verify the potential literary data mining applications

In this stage we are going to interview 5-10 literary scholars to examine 1) how the scholars discover the kinds of knowledge identified in stage 2; 2) how to specify these kinds of knowledge so that computational algorithms can be designed to discover them for literary study purpose.

KDD Title Words	KDD TF	MUSE DF	MUSE TF	MUSE DF-pcnt	MUSE TF-ppm	ANC-NYTIMES DF	ANC-NYTIMES TF	ANC-NYTIMES DF-pcnt	ANC-NYTIIMES TF-ppm
cluster	52	13	17	10	22	50	56	1	24
model	40	99	438	80	562	335	597	8	254
pattern	29	49	121	40	155	167	207	4	88
Net-work	23	30	76	24	97	325	724	8	307
classif classifi	35	14 19	64	11 15	82	16 50	74	0 1	32
rule	19	81	210	65	269	708	1409	17	598
associ	15	103	479	83	614	762	1161	18	493
graph graphic	15	2 15	25	2 12	32	7 244	504	0 6	214
stream	15	19	26	15	33	103	117	2	50
serial seri	10	20 80	213	16 65	273	32 537	94	61 13	403
relat relation-ship	10	117 98	1367	94:79	1753	386 323	911	9 8	487
framework	10	38	69	31	88	25	28	1	12
correl	9	20	30	16	38	15	21	0	9
similar similarli	9	99 66	475	53 80	609	484 77	649	12 2	276
spatial	7	19	55	15	71	8	8	0	3
decis	7	57	151	46	206	683	1110	16	471
hierar-ch(i)	6	20 41	178	16 33	229	4 4	45	0 0	13
sequen(c/ti)	6	34 8	80	6 27	103	4 51	71	0 1	30

Table 1: KDD keyword frequency comparison between MUSE and ANC-NYTIMES

Note: Because of the limitation of Porter Stemmer, some words with the same stems have to be manually merged together, such as “graphs” and “graphics”. In these cases the TF-ppm can be summed up, but the DF-pcnt can not be merged, so both DF-pcnts are listed.

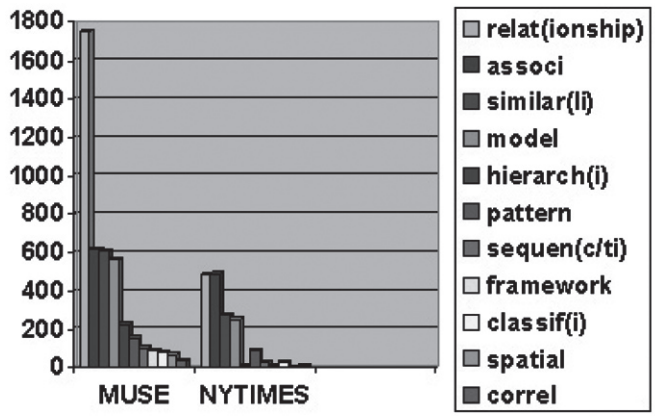


Figure 1: The frequencies (in ppm) of KDD keywords in MUSE and NYTIMES

References

Buckland, M. (1999). Vocabulary as a central concept in library and information science. In Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities, *proceedings of the Third International Conference on Conceptions of Library and Information Science*. Dubrovnik, Croatia, pp. 3–12.

Mei, Q. and Zhai, C. (2005). Discovering evolutionary theme patterns from text. *Proceedings of The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Chicago, Illinois, pp. 198-207.

Meunier, J. G., Forest, D. and Biskri, I. (2005). Classification and categorization in computer-assisted reading and text analysis. In Cohen, H. and Lefebvre, C. (eds), *Handbook on Categorization in Cognitive Science*. Elsivier.

Unsworth, J. (2000). Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this? *Symposium on Humanities Computing: formal methods, experimental practice*. King’s College, London. Available: http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html

Multiple Sessions

AN ODD BASKET OF ODDs

Syd BAUMAN

Women Writers Project, Brown University

The Text Encoding Initiative's *Guidelines for Electronic Text Encoding and Interchange* version P4 provides for separate steps of 'customization' and 'extension'. The former consists of choosing which TEI tag-sets to use, and is typically performed either in the document type declaration subset or using the PizzaChef. The latter consists of various other alterations, including renaming elements, adding elements, deleting elements, and modifying content models or attribute lists. Although the PizzaChef can help ease this process, most of the work is done by re-defining TEI parameter entities using the DTD language. An 'extension' may extend the set of documents valid against the resulting DTD, may restrict that set, or may create a completely non-overlapping set. In all cases, it is still called an 'extension'.

In TEI P5, these two steps have been rolled into one. The selection of modules (including the few mandatory ones) is performed in a customization file. Changes to the schema, including the deletion of elements, the addition of elements, etc., are also performed in this same customization file.

Because module selection is a necessary process, in order to use P5 a user must start with a customization file (even if it is just one of the sample customization files shipped with TEI). Luckily, the P5 method of creating customizations is substantially simpler than P4's DTD method. This means they are easier to write, easier to document, and easier to share.

P5 customization files are written in TEI, using the module for tagset documentation. This is also true for the Guidelines themselves — they are written in TEI, using the module for tagset documentation. Because using this tagset one encodes a single file from which all the needed outputs are produced (the prose Guidelines, the reference documentation, and the schemata), this language is called ODD, for "one document does it all". Therefore the customization files are usually referred to as "ODD files", and given the extension .odd.

Because ODD files are written using a language the user already knows (TEI), rather than a special schema-language syntax, creation of customizations is quite a bit easier. Furthermore, the TEI class system (which groups elements for convenient use) and datatype system (which provides restrictions for attribute values) are significantly simplified since P4. And last, but far from least, there is a significant web interface ("Roma" at http://www.tei-c.org.uk/Roma/) that presents a form interface for creation of the ODD customization files. This means that instead of trying to remember whether the correct identifier for the module for names and dates is "namesdates", "namesAndDates", "namesDates", etc., the user need only click on the "add" next to "Names and Dates".

Because the same language as that used for the Guidelines themselves is used, brief documentation about added or changed elements or attributes can be placed directly within the customizations themselves. Examples of usage and remarks can also be encoded within the customizations, and standard TEI style-sheets will produce reference documentation from these. Furthermore, because ODD files are written in TEI, the usual prose elements (like <p>) can be used alongside the specialized customization elements to provide more in-depth documentation for the customization in the same file. finally, since a customization file is a TEI XML file, it can be validated and easily transferred via any Unicode aware exchange method. Because it contains the documentation as well as the customizations themselves, it can be easily understood by the recipient.

But what do customizations look like? What are they good for? How does one make them? What are the advantages and disadvantages of various customizations? This panel presentation hopes to answer these questions not by having TEI "experts" lecture, but rather by having actual TEI users who have performed P5 customizations present their projects.

Each of the following projects will give a brief presentation including a discussion of the reasons and motivations behind their customization, a careful look at the ODD file itself, a glance at the output reference documentation (some of which will be non-English), and perhaps an overview of the tool-chain used, culminating in a demonstration of sample output.

The Henry III Fine Rolls Project

Paul SPENCE
Centre for Computing in the Humanities, King's College London

The Henry III fine Rolls Project is a three year project which aims to publish a series of legal documents, known as 'fine rolls', which chart offers of money made to King Henry III of England in exchange for a wide range of concessions and favours. There is a fine roll for each of his fifty-six years on the throne (1216–1272) and they offer significant insight into the political, governmental, legal, social and economic history of the day. Based at King's College London and the UK's National Archives, the project aims to publish the rolls up to the year 1248 in both book and electronic form.

The electronic version will enable the indexing and searching of a wide variety of legal terms relating to office, finance, witnesses and the nature of the legal documents themselves, as well as more general terms relevant to the period, such as those associated with social status, ethnicity and family structure.

Our use of TEI inevitably involves a large degree of customisation: partly due to the specific legal nature of the fine rolls (hence the addition of new elements), partly due to the practical needs inherent in any multi-editor collaborative environment (which have made us examine ways to simplify markup structures) and partly due to the rich 'subject' terms which are being encoded (leading to our tailoring the attribute values permitted for a given term). I will assess the impact on this project of the new customisation facilities available in P5, describing the customisation itself, the consequences for work-flow in an ever-evolving scholarly process, the production of documentation, and the wider benefits for the publication process.

Tagalog Lexicon

Michael BEDDOW
Independent Scholar

The encoding of lexicographical data for Tagalog and related languages presents significant challenges. While the P4 additional tagset for print dictionaries (and the current instantiation of the P5 module) has proved quite adequate for encoding dictionaries of Western languages, the morphology and semantic patterning of languages in the Philippine sub-group of the West Austronesian family require extensions which provide a good indicator of the power and flexibility of P5 customization techniques.

The Mapas Project

Stephanie WOOD, Judith MUSICK
Wired Humanities Project, University of Oregon

The Mapas Project is a multi-year project that aims to publish exemplary early Mesoamerican pictorial manuscripts in electronic form and indexed to allow for scholars working on these or similar manuscripts to find key terminology or pictorial elements from multiple manuscripts to support philological, historical, art historical, anthropological, geographical, and linguistic research.

We will start by showing and describing one particularly interesting manuscript, the Mapa de Mixtepec. This sizable manuscript (47 cm x 74 cm) was painted on deer hide in the late seventeenth century. It comes from a Zapotec town (San Andrés Mixtepec) in what is now the state of Oaxaca. It shows the people's ancient origins in Zaachila, their historic and legendary migration, the founding of their town, a genealogy of their leading families, and their colonization by Christians. The mapa is primarily pictorial with multiple textual elements arranged and relating to one another cartographically rather than in any particular linear order.

Our current aim is to encode the secondary textual material developed by Wood and other scholars in TEI P5, directly linking various annotations to the image, and describing the relationship of the annotation to the indicated section of the image. With the possible exception of problems posed by special terminology such as selecting restricted sets of attribute values to encode key terms in indigenous languages (e.g., "altepetl", the term for the key Nahua socio-political unit), the adaptations of TEI will be fairly straightforward. There are additional complications, however, including the aforementioned need to create relationships between annotations and the whole document, relating text and image to each other and to their actual location within the manuscript, and

due to the special nature of the manuscripts and the multi-editorial and internationally collaborative nature of the project

Resources for Research on Tang Civilization

Christian WITTERN
Institute For Research In Humanities, Kyoto University

The Project "Resources for Research on Tang Civilization" attempts to provide a wide range of original texts in Chinese pertaining to the research of the Tang period (7th to 10th century). This includes histories, annals and administrative documents, but also literary, religious and philosophical works, compilations and encyclopedias. The networked workspace we plan to provide will allow the researcher to navigate in the multifaceted and multidimensional conceptual space and enable him to record his discoveries, and thus contribute to the resource while using it.

We use TEI at different levels in this project. In a heavily customized version, we capture the texts and our affirmations about them. The customizations mainly reflect constructs necessary for certain features of Chinese texts, but are also partly due to special requirements of our workflow. While we started out using a hand-crafted version derived from a P4 DTD, we later converted our customization to P5 using ODD. In this presentation, I will introduce this customization and talk about the process of developing and applying them.

Epigraphic Documents in TEI XML (EpiDoc)

Zaneta AU
King's College London

Gabriel BODARD, Tom ELLIOTT

Under the rubric of 'EpiDoc' one finds a growing community of scholars intent upon using the TEI for the digital encoding and interchange of epigraphic documents; i.e., modern digital editions—both print-derived and 'born digital'—of inscribed and incised texts emanating from the areas associated with the Roman empire at any period before the medieval. This community's chief deliverable to date has been a set of guidelines setting out agreed standards for the application of TEI P4 to such work, together with a customized DTD and various software tools. The success of the effort is demonstrated, and has been advanced by, three flagship projects: one at King's College London under the direction of Charlotte Roueché (Aphrodisias in Late Antiquity (2d. ed.) and The Inscriptions of Aphrodisias) another at Brown University under the direction of John Bodel (The U.S. Epigraphy Project) and a third at Oxford University under the direction of Alan Bowman, Charles Crowther, and John Pearce (The Vindolanda Tablets Online).

The EpiDoc community is poised to move from P4 to P5. At P4, our extensive and growing guidelines were developed under the auspices of the same customized DTD that we promulgated for epigraphic encoding, with various resulting infelicities and shortcomings. Our strategy for P5 is to separate the schemata for the EpiDoc Guidelines from that for the encoding recommended by those guidelines. In moving the EpiDoc Guidelines to P5 first, we create a standard, but highly structured, TEI ODD that will, in turn, instantiate a P5 schema to govern EpiDoc-compliant encoding practice.

We shall begin with a brief overview of the EpiDoc Guidelines, focusing on those features that are dictated either by the requirements of epigraphic markup itself or by the community's open development practices, in which software tools and the EpiDoc Guidelines themselves are multi-authored by individuals whose TEI expertise varies considerably one to another. Given this mode of collaboration, the flexible application of constraint possible with the RelaxNG schema technology has significant advantages over the Backus-Naur content models of P4. We shall also discuss the elements we have added as extensions or adaptations to the P5 tagset, such as tags for typographic forms and those supporting the integration of linked typographic examples, encoding examples and regular expressions. These data 'tuples' form the basis for automated verification of EpiDoc-specific XSLT stylesheets and custom text conversion tools, as well as the internal consistency of the EpiDoc Guidelines themselves.

QUANTITATIVE CODICOLOGY AND THE REPERTORIUM WORKSTATION

David BIRNBAUM

Department of Slavic Languages, University of Pittsburgh, USA .

The Repertorium of Old Bulgarian Literature and Letters (henceforth "the Repertorium") is a series of projects begun in 1994 by Anisava Miltenova (Institute of Literature, Bulgarian Academy of Sciences), David J. Birnbaum (Department of Slavic Languages and Literatures, University of Pittsburgh), Andrej Bojadzhiev (University of Sofia), and Milena Dobreva (Institute of Mathematics and Informatics, Bulgarian Academy of Sciences) and designed to explore the opportunities for exploiting computational resources to advance the study of Slavic philology in original and innovative ways. From its inception the Repertorium set out to transcend the impressive but ultimately limited achievements of many other projects in humanities computing, such as structured document development, structured searching, variable rendering for multiple uses, etc. While the Repertorium has appreciated these technologies and has used them productively, its long-term goal has been not merely to enhance traditional philological activities (such as publishing and searching), but also to facilitate new types of philological research that would have been not merely impractical, but essentially unthinkable without the use of computational tools.

Over the past ten years the Repertorium has overseen the encoding of more than 350 manuscript descriptions by members of the Institute of Literature, and it has also coordinated and directed compatible large-scale joint projects with British and Swedish researchers. The Repertorium DTD, begun before both the TEI P5 manuscript description module and the European MASTER initiative, is designed as a conformant modification of the TEI DTDs, but it imposes tighter and more consistent structure than any comparable project of which we are aware, and this attention to structure (whereby a manuscript description is as much a database as a document) is designed specifically to facilitate not merely rendering or structured searching, but also data-mining for analytical purposes, an enterprise for which I have coined the term "quantitative codicology."

In a keynote address at one of the recent Extreme Markup conferences, Tommie Usdin (Mulberry Technologies) observed that "XML has made true all of the lies that we told about SGML." When the Repertorium team began encoding manuscript descriptions in SGML the mid-1990s, we took on faith that at some point it would be possible to perform the sort of analysis we needed for our data, but it is only with the introduction of such relatively recent ancillary XML technologies as XSLT, XQuery, and SVG that it has been possible to exploit the Repertorium inventory for the innovative and sophisticated analysis that is was originally intended to facilitate.

My presentation at the 2006 Digital Humanities conference will concentrate on describing both the technologies developed within the framework of the Repertorium initiative for analyzing medieval Slavic manuscript materials and the philological principles that guided our research. In particular, this presentation will demonstrate the ability for a scholar seated behind a standards-conformant web browser anywhere on the Internet to combine on the fly and in varied ways structured searching with the generation of such graphic representations of the relationships among manuscripts as dendrograms and plectograms (about which see my "Computer-Assisted Analysis and Study of the Structure of Mixed-Content Miscellanies," Scripta & e-Scripta,. vol. 1 (2003), 15–54 [preprint available on line at http://clover.slavic.pitt.edu/~djb/2003_ljubljana/2003_ljubljana_paper.pdf]).

What is most new and significant since my initial introduction of these technologies into manuscript studies is my integration of them and of other tools into a "Repertorium Workstation," an integrated and coherent platform where users can browse and search manuscript descriptions, generate plectograms on the fly (through behind-the-scenes XQuery and XSLT transformations to SVG), use the plectograms to launch new queries, and, in general, explore the range of relationships among manuscript witnesses without constraint in ways that have never been exploited before in humanities

computing, and that were essentially unthinkable due to technological limitations until very recently. The presentation will combine a demonstration of the Repertorium Workstation with a discussion of its philological motivation, its design principles, the types of original and innovative primary scholarship that it supports, and plans for further development.

All Repertorium materials, from the raw data files to the Internet-based Workstation, are (or will be) freely accessible on the Repertorium web site (http://clover.slavic.pitt.edu/~repertorium/). While many of the technologies underlying the Repertorium Workstation were developed to address specific needs of Slavic medievalists (e.g., Unicode-based support for multiple and uncommon writing systems), the system has great general applicability to medieval studies in other cultural tradition, as well as to non-medieval studies of textual traditions. In particular, its innovative use of SVG to model the structural similarities among manuscript witnesses can serve as an example (easily adapted to other projects) of how graphic representations can provide insights into textual information that would otherwise remain virtually imperceptible.

Metadata and Electronic Catalogues: Multilingual Resources for Scientific Medieval Terminology

Andrej BOJADZHIEV
University of Sofia

The Metadata and Electronic Catalogues project (2004–06), a component of the Repertorium of Old Bulgarian Literature and Letters, is designed to create electronic catalogues and authority files that will serve as integrated repositories of terminological information that has been developed and applied successfully in already existing projects in the realm of medieval Slavic languages, literatures, and cultures. One innovative feature of this project is that beyond serving as a central repository for such information, it will expand the organizational framework to support a multilingual superstructure along the lines of I18N initiatives elsewhere in the world of electronic text technology in general and humanities computing in particular.

This project is based on distinguishing the meanings of particular terms and notions that appear in the text of medieval written documents both within the primary corpus and in comparison to established scholarly terminology (for example, medieval Slavic writers used several terms—sometimes systematically and sometimes not—to identify what modern scholars might call, variously, a sermon, a homily, an instruction, etc.). This orientation is designed to support the development and implementation of software tools for finding multilingual counterparts to both original (medieval) and scholarly (modern) terminology, and for sorting, searching, and mining this data in a way that is independent of both the entry and the query languages.

The project presumes the possibility (unique for such initiatives in Slavic humanities computing, and uncommon in humanities computing in general) of linking the standardized terminological apparatus for description, study, edition, and translation of medieval texts, on the one hand, to authoritative lists of key-words and terms used in bibliographic descriptions, on the other. This will allow the integration of scholarly meta-data and bibliographic references under a single unified framework. Another aim of the project is to create a mechanism for allowing the extraction of different types of indices based upon the imported documents even when the languages of encoding may vary, so that, for example, Serbian-language descriptions of medieval Slavic (not necessarily Serbian) manuscripts could be collated with Bulgarian-language descriptions of other medieval Slavic manuscripts in a way that would enable automated systems (such as the plectogram-generating tool described in David J. Birnbaum's proposal for our panel) to recognize when entries that differ textually are nonetheless to be treated as the same. The primary manuscript description texts are encoded in a TEI-based XML format in the context of the broader Repertorium initiative, and their utility for the type of multilingual authority files, bibliographic databases, and other broad reference resources illustrates the multipurposing that is characteristic of XML documents in the humanities, but on a broader scale than is usually envisioned (that is, going beyond the more common tasks of transformation into multiple presentation formats or exploitation in structured searching).

The Metadata and Electronic Catalogues project is

based at the Institute of Literature, Bulgarian Academy of Sciences, where the working team concists of Anisava Miltenova (Institute of Literature, director of the project), Ana Stojkova (Institute of Literature), Andrej Bojadzhiev (Sofia University), Margaret Dimitrova (Sofia University), and Svetla Koeva (Institute of Bulgarian Language BAS).

The outcome of the project is twofold:

1. The development of a database that will provide the terminological and bibliographic apparatus that will support the study of the medieval Slavic texts for both research and educational purposes.
2. The development of an on-line query interface that will enable this database also to serve as a sort of independent encyclopedic reference to medieval Slavic written culture.

The Repertorium Intitiative: Computer Processing of Medieval Manuscripts

Anissava Miltenova
Institute of Literature, Bulgarian Academy of Sciences

The application of computer technologies to store, publish and—most importantly—investigate written sources belongs to the most promising tasks at the boundary between the technical sciences and the humanities. The Repertorium Initiative was founded in 1994 at the Department of Old Bulgarian Literature of the Institute of Literature of the Bulgarian Academy of Sciences in collaboration with the University of Pittsburgh (US). The Repertorium is a universal database that incorporates archeographic, paleographic, codicological, textological, and literary-historical data concerning the original and translated medieval texts distributed through Slavic manuscripts between the eleventh and the seventeenth centuries. These data include both parts of actual texts and the results of their scientific investigation, with particular attention to the study manuscripts typology, a traditional aspect of philological scholarship that has been reinvigorated by the introduction, through the Repertorium Initiative, of computational methodologies.

The descriptions and examples of real texts are based on the XML (Extensive Markup Language), an informatic standard that incorporates special "markup" characters within natural language texts. The markup tags demarcate certain parts of the texts (elements) and signal what the data represents, simplifying the identification and extraction of data from the text not just during conversion for rendering (the most common procedure in humanities projects), but also during data-mining for analysis. The most recent model of description of manuscripts in an XML format derived from the TEI (Text Encoding Initiative) guidelines has been developed by Andrei Bojadzhiev (Sofia University), following five main principles, formulated in the context of the project by David J. Birnbaum in 1994:

1. Standardizing of document file formats;
2. Multiple use (data should be separated from processing);
3. Portability of electronic texts (independence of local platforms);
4. Necessity of preservation of manuscripts in electronic form;
5. The well-structured division of data according to contemporary achievements in textology, paleography and codicology.

he working team in the Institute of Literature has already developed a digital library of over 350 electronic documents. Since its inception as a joint Bulgarian-US project over ten years ago, the Repertorium Initiative has expanded to include a joint Bulgarian-British project describing Slavic manuscripts in the collection of the British Library (London), as well as a project with University of Gothenburg (Sweden) concerning the study of late medieval Slavic manuscripts with computer tools. The Repertorium Initiative has grown not only in terms of its geography and its participants; it has also come to include a unique set of possibilities for linking the primary data to a standardized terminological apparatus for the description, study, edition, and translation of medieval texts, as well as to key words and terms used in the bibliographic descriptions. This combination of structured descriptions of primary sources with a sophisticated network of descriptive materials permits, for example, the extraction of different types of indices that go well beyond traditional field-based querying.

In recognition of the ground-breaking achievements of The Repertorium Initiative, its directors and principal researchers were appointed in 1998 by the International Committee of Slavists (the most important such

international association) to head a Special Commission for the Computer Processing of Slavic Manuscripts and Early Printed Books. Other evidence of the achievements of this project include, the organization three international conferences (Blagoevgrad 1994, Pomorie 2002, Sofia 2005) and the publication by the Bulgarian Academy of Sciences of three anthologies (1995, 2000, 2003). The Internet presence of the Repertorium Initiative is located at http://clover.slavic.pitt.edu/~repertorium/ .

Because the Repertorium Initiative goes beyond manuscript studies in seeking to provide a broad and encyclopedic source of information about the Slavic medieval heritage, it also incorporates such auxiliary materials as bibliographic information and other authority files. In this capacity the Repertorium Initiative is closely coordinated with three other projects: the project for Authority Files, which defines the terms and ontology necessary for medieval Slavic manuscript studies; Libri Slavici, a joint undertaking of the Bulgarian Academy of Sciences and the University of Sofia in the field of bibliography on medieval written heritage; and identifying the typology of the content of manuscripts and texts with the aid of computational tools. All three of these share the common structure of the TEI documents and use a common XSLT (Extensible Stylesheet Language for Transformations) library for transforming documents to a variety of formats (including XML, HTML [Hypertext Markup Language], and SVG [Scalable Vector Graphics]) thus providing a sound base for the exchange of information and for electronic publishing.

The relationship among the three projects could be described in the following way:

1. The Repertorium Initiative is a innovative from both philological and technological perspectives in its approach to the description and edition of medieval texts. It takes its metadata for description from its Authority Files and its bibliographic references from the Libri Slavici.
2. The Authority Files project gathers its preliminary information on the basis of descriptions and prepares guidelines in the form of authority lists for the use the metadata by researchers.
3. Libri Slavici accumulates its data from various sources, including descriptions and authority files, and shares common metadata with both of them.
4. Visualization of typology is radically new non-textual representations of manuscript structures. This development demonstrates that computers have done more than provide a new way of performing such traditional tasks as producing manuscript descriptions. Rather, the production of electronic manuscript descriptions has enabled new and innovative philological perspectives on the data.

The future of the Repertorium Initiative is to continue integrate into a network full text databases of medieval Slavic manuscripts, electronic description of codices, and electronic reference books with terminology. Preserving the cultural heritage of European libraries and archives, it provides for data and metadata search and retrieval on the basis of paleographic, linguistic, textologic, and historical and other cultural characteristics. The connections among the different subprojects thus lead to a digital library that is suitable for the use of a wide community of specialists, and, in the same time, continues to inspire related new projects and initiatives.

References

Miltenova, A., Boyadzhiev, A. (2000). “An Electronic Repertory of Medieval Slavic Literature and Letters: a Suite of Guidelines”. In: *Medieval Salvic Manuscripts ans SGML: Problems and Perspectives*. Sofia:“Marin Drinov” publishing house, 44-68.

Miltenova, A. Boyadzhiev, A. Radoslavova, D. (2003). “A Unified Model for the description of the Medieval Manuscripts?”. In: *Computational Approaches to the study of Early and Modern Slavic Languages and Texts*. Sofia:“Boyan Penev” Publishing Center, 113-135.

THE RHETORIC OF PERFORMATIVE MARKUP

Julia FLANDERS

Women Writers' Project, Brown University

In the classic account first proposed by DeRose et al. (1990) and subsequently developed by Renear et al. (1996) and finally by Renear (2001), text markup of the sort typically practiced by humanities computing scholars is a reflection of reality. It seeks to express observations about the nature of the text, rather than giving orders to a processor. Markup of this sort—labeled as "logical indicative" markup in the most recent formulation—is widely familiar in the scholarly community, instantiated in markup languages like TEI, EpiDoc, and other languages intended for the transcription and preservation of primary sources.

What we think of as logical indicative markup, however, is almost never that simple. Although markup in the indicative mood claims to advance a simple statement of fact ("This is a paragraph"), the actual intellectual activity being undertaken is often much more complex. Despite the early claims of the OHCO theorists that markup observes what text "really is", it is only under tightly constrained circumstances that indicative markup truly makes something approaching factual observations concerning textual features. Within a given disciplinary community the identification of a paragraph or a line of verse may be uncontroversial (so that one would appear to be quibbling if one paraphrased as "I believe this to be a paragraph"). Within a digital library context the encoding may be so slight that its claims about the text carry almost no information (so that marking something with may mean only "this is a block of text") and hence no information with which it would be possible to disagree. But if we broaden the context at all—using markup to communicate between disciplinary groups, or to describe more complex documents—we enter a very different and less factual terrain. For any early work where the modern generic distinctions are not yet solid, identifying a passage of text as, variously, a paragraph, a verse line, an epigraph, or some other less determinate segment is not an act of factual observation and correct identification, but of strategic choice. The question is "how does it make sense to describe this textual feature?" rather than "what is this feature?" And implicit in the idea of "making sense" are qualifiers such as "for me", "now", "for my present purposes", "here at this project", "given my constraints", and others that can readily be imagined. The choice of the phrase "making sense" is not casual here: the act of encoding is indeed an act of making sense, creating conditions of intelligibility.

Renear, in his essay "The Descriptive/Procedural Distinction is Flawed," extends the earlier taxonomy of markup types by adding the dimension of "mood", by which markup may be characterized as indicative, imperative, or performative. Where indicative markup is the kind described above—making factual statements about the textual world—and imperative markup is the kind that issues a command (for instance, to formatting software), performative markup is a less familiar domain, which Renear identifies with authoring. As his phrase "markup that creates" suggests, this domain has to do with calling text into being and in particular with naming and effecting the structures through which that text expresses meaning. It is tempting to make a clean distinction between this kind of "authorial markup" and the more familiar indicative markup on the basis of the type of document concerned: authorial, performative markup being what we use when we write new documents, and ordinary indicative (or perhaps "editorial") markup being what we use when we transcribe existing source material. However, having noted this distinction we must immediately trouble it: first, because these categories are often intermingled (for instance, annotations and commentary in a scholarly edition are "authored" in this sense). But more significantly, even with content that is not "new", markup does not exist solely to name what is there, but also serves to express views about it, and the expression of these views constitutes an authorial act just as surely as the generation of a sentence of commentary. Authorial markup brings a structure into being just as writing brings words into being, and in some cases the two may be isomorphic. Adding the TEI element amounts to the same thing, informationally speaking, as adding a note whose content is "this sentence is unclear in the source; we believe the reading to be X." To extend Renear's terms, can be either an indicative, editorial statement ("this passage is unclear") or a performative, authorial statement ("I make this assertion about unclarity", "I create this

meaning with respect to the unclear reading").

This authorial dimension to markup systems like the TEI is unfamiliar, little used, obscure. But it crucially amplifies our understanding of the rhetoric of markup, and of the kinds of meaning it can carry. Most importantly, it suggests that Jerome McGann's assertion that text markup cannot represent "the autopoietic operations of textual fields—operations specifically pertinent to the texts that interest humanities scholars" (2004, 202-3) reflects a very limited sense of the potential rhetorical operation of markup. In characterizing the TEI as "an allopoietic system" which "defines what it marks...as objective", McGann draws on markup's own conventionalized account of itself. This account, which as we have seen locates systems like the TEI firmly within the indicative realm, deals solely with the editorial rhetoric of statement and description—not with interpretation and certainly not with performance. It ignores the extent to which even this indicative markup can make statements which are not simply factual: which represent local knowledge, perspective, contingency, belief, positionality, uncertainty, purposiveness, and even deception.

Most importantly, it ignores the authorial quadrant of Renear's grid: the space of performative logical markup, in which an author brings meaning into existence either by creating new marked content, or by adding markup to an existing text and performing upon it a new set of meanings. This latter case would in fact resemble performative instruments like the Ivanhoe Game, which represent for McGann the archetypal scholarly textual activity: a performative apparatus, in effect, through which scholars express interventions in a textual field: "readings", commentary, textual engagements that inflect the object text rather than simply standing apart from it. The extended version of this paper will expand on this point, exploring how performative or authorial markup might enact the kinds of textual engagements that McGann calls for as constitutive of humanistic textual study.

McGann is correct in identifying the predominant use of markup systems like TEI as "coding systems for storing and accessing records" (202). But this predominant use does not define the limits of capability for such markup systems, let alone for text markup in general. Our choice to use markup in this way derives from the collective sense, within the humanities disciplines, that this is what markup should be for: a technological tool external to ourselves, rather than an expressive medium. In expecting, as McGann does, that markup cannot be made of the same stuff as poetry, we create a self-fulfilling prophecy. In fact, we make markup in our own image—in the image of our own fears.

References

DeRose, Steven, J., David Durand, Elli Mylonas and Allen H. Renear. "What is Text, Really?," *Journal of Computing in Higher Education*. 1:2 (1990).

Liu, Alan. "Transcendental Data: Toward a Cultural History and Aesthetics of the New Encoded Discourse." *Critical Inquiry* 31:1 (Autumn 2004), 49-84.

McGann, Jerome. "Marking Texts of Many Dimensions" in *A Companion to Digital Humanities,* ed. Susan Schreibman, Ray Siemens, and John Unsworth. Blackwells, 2004.

Renear, Allen. "The Descriptive/Procedural Distinction is Flawed." Markup Languages: T*heory and Practice* 2.4 (2001): 411-420.

Renear, Allen, Elli Mylonas, and David Durand. "Refining our Notion of What Text Really Is: *The Problem of Overlapping Hierarchies." Research in Humanities Computing,* ed. Nancy Ide and Susan Hockey. Oxford University Press, 1996.

The Rhetoric of Digital Structure

Clifford WULFMAN
Modern Culture and Media, Brown University

This paper will examine the rhetoric of textual markup by relating it to mapping. I begin with the observation that markup is a kind of poesis, which is itself a species of mapping: to paraphrase Theseus at the end of A Midsummer Night's Dream, it is the embodiment of form, the giving to airy nothing a local habitation and a name. I will discuss so-called "mapping art," a digital

art form that uses data sources, or streams, or pools, or bases, and filters them to create works of art that are, in some way, connected to the source, if only functionally. I will focus on pieces that claim informational status and discuss the way rhetoric inevitably inflects the transformation/mapping of data into information. These pieces, I will claim, shed a new light on the rhetorical nature of textual markup.

This by itself is nothing new: we've long understood that all marking is interpretation. But I will extend this observation to discuss the relation of mapping and marking to information and reading. I will examine linguist Geoffrey Nunberg's discussion of the term "information," comparing it with Shakespeare's notion of poeis as (ex)formation, and consider the proposition that reading is information, and there can be no information without a map.

The argument turns at this point to consider the possibility that markup is not mapping so much as it is tracing: not the translation of data from one domain to another but a kind of delineation, a marking on the body of the text itself. Thus "informing the corpus" is replaced with "inscribing the corpse," and instead of Shakespeare's depiction of imaginative creation, we have Kafka's fascist nightmare In the Penal Colony.

Between these two poles must be a middle way of reading, and I will conclude by briefly considering how the Lacanian notion of the Phallus can help us understand the desire for markup, with a gesture towards Harold Bloom's anxieties of Influence and maps of misreading.

The Rhetoric of Mapping Interface and Data

Elli MYLONAS
Brown University

In this paper I want to continue to the discussion of the rhetoric of the digital monograph I presented at ACH-ALLC2005. That paper looked at past discussions of rhetoric as applied to scholarly hypertext and to the web.

As a continuation of this line of thought, I'd like to consider the relationship of underlying "data" to user-facing "interface." Two significant current theoretical models of digital publication are the notions of the "digital archive" and the "database." Theorists like Daalgard and Moulthrop see the internet as a global archive; their choice of term privileges the function of the collection of data, its completeness and lack of a singular perspective or notion of truth. For them, the archive is specifically a very large collection of homogeneous or heterogeneous documents available in digital form, like the internet, or the NY Times archive. Lev Manovich uses the term "database" as the informing paradigm for the organization of new media productions; his terminology privileges the mode of interaction enables by a particular technology. For Manovich the "database" is a more suitable replacement for "narrative" when discussing the potential and effect of digital art and other (constrained) websites. Like Moulthrop and Daalgard, Alan Liu uses the term "archive" for digital collections that are XML-based, as opposed to scholarly collections that are database driven, and views both sorts of collection as functionally equivalent.

In this paper, I'd like to move between the larger and more generic archive and the database informed individual website. Manovich's idiosyncratic use of the term database to refer to any structured data can be confusing, but in this paper, implementation methods, whether XML markup, database, or some other technology, will not affect the discussion.

Alan Liu has described the change in user interface and in the role of the scholar/author as more websites become template driven. He identifies a shift from the craftsmanlike activity of early website creation, to a model where the interface and the underlying source data are intentionally separate, as are the roles of their collector and designer. Interface designers are designing pages with empty space into which data pours, over the content of which they have no control. The user interface is usually considered the locus of meaning in a website or in a digital publication. In the case of the contained, authored "monograph" the user interface is where an author can present a point of view and engender behaviors in the readers of the work. At first glance, even in a digital archive, the user interface determines what a user can do, or learn about the underlying material. However, the underlying data, which may be structured either as a database or using XML markup, also is inextricably linked to the interface. Rhetorical casts that have been inserted into the data interact with the

interface, just as the interface affects not only users' technical ability to manipulate the data, but their view of what manipulations make sense for that data.

This interaction occurs in the space between the data and the interface. That space contains the process enabled by a web site whose most important part can be described as a mapping between interface and data. This mapping necessitates an interaction of surface and infrastructure and can only be understood in the context of an awareness of the rhetoric that belongs to each side of the map. Part of this space is the domain of information designers and interface designers. They are the ones who plan user interactions, lay out the relationships of pages that a user sees, and are familiar with human cognitive ability and usability. But before one can draw the map of the presentation, the map of the content must exist, and this is also an intentional product. Because of this, at the most significant level, these decisions inhere in the scholar who is amassing the archive or who is authoring the monograph. The scholar can indicate what to markup, and what kinds of interactions a user should be able to have with the digital work.

The importance of the relationships between "source" and "visualization" or "content" and "presentation" has already given rise to a new genre of art, that explores these relationships in a playful way. Artists represent existing, often real-time, data streams such as internet traffic, economic or geophysical data using visual representations. Manovich suggests that one way to engage critically with such works is to look at how effectively the choice of mapping functions as a commentary on the data stream. This same approach may be applied to digital publications in order to evaluate and understand the dependencies between interface and data.

These theories will be tested by discussing some representative websites such as the Women Writers Project (www.wwp.brown.edu), documents in the Virtual Humanites Lab (http://www.brown.edu/Departments/Italian_Studies/vhl/vhl.html) and Thomas and Ayers, The Differences Slavery Made (http://www.vcdh.virginia.edu/AHR/).

References

[1] **Daalgard, Rune**, "Hypertext and the Scholarly Archive - Intertexts, Paratexts and Metatexts at Work", in *Proceedings of the twelfth ACM conference on Hypertext and Hypermedia* (14-18 August, Aarhus, Denmark). New York: ACM Press: 175-184

[2] **Liu, Alan**, "The Art of Extraction: Toward a Cultural History and Aesthetics of XML and Database-Driven Web Sites." http://dc-mrg.english.ucsb.edu/conference/2002/documents/Alan_Liu_Art_of_Extraction.html.

[3] Transcendental Data: *Toward a Cultural History and Aesthetics of the New Encoded Discourse*. By: Liu, Alan. Critical Inquiry, Autumn 2004, Vol. 31 Issue 1, p49-84

[4] **Manovich, Lev**, T*he Language of New Media*. MIT Press, 2002.

[5] **Manovich, Lev**, *The Anti-Sublime Ideal in New Media.* Chair et Metal 7, 2002. http://www.chairetmetal.com/cm07/manovich-complet.htm

[6] **Moulthrop, S**. *The analog experience of digital culture, in R. Koskimaa and M. Eskelinen* (eds.), Cybertext Yearbook 2000. Jyvaskyla: Publications of the Research Centre for Contemporary Culture, 2001. pp. 183-98.

[7] **Moulthrop, Stuart**, "What the Geeks Know: Hypertext and the Problem of Literacy" in *Proceedings of Sixteenth ACM Conference on Hypertext and Hypermedia,* ACM Press 2005, pp xx-yy.

[8] **Christiane Paul**: Databases, data visualization and mapping in Christiane Paul: *Digital Art, Thames and Hudson* 2003, pp. 174-189

[9] **Ryan, Marie-Laure**, *Cyberspace, Cybertexts, Cybermaps. Dichtung-Digital* 2004, issue 1. http://www.dichtung-digital.org/2004/1-Ryan.htm

THE NORA PROJECT: TEXT MINING AND LITERARY INTERPRETATION

Matthew KIRSCHENBAUM
Panel ABSTRACT

This panel brings together three papers showcasing different facets of the nora Project, a multi-institutional, multi-disciplinary Mellon-funded initiative to apply text mining and visualization techniques to digital humanities text collections.

We are currently one year into the initial two-year phase of the project. Though most of our methods remain tentative, most our findings speculative, and our technical environment experimental, we nonetheless have significant progress to report. In practical terms, work on the project has advanced considerably since the initial demos and research agendas that were presented at last year's conference (2005). We have conducted four sustained text mining investigations (two of which are discussed in detail in the papers below), built a complete technical environment that allows a non-specialist user to engage in the text mining process, and we have begun to achieve some consistency in our understanding of what data mining in the humanities, particularly literary interpretation, might be good for. While our findings in this last area remain contingent in the extreme, they nonetheless tend to cluster around activities such as provocation, patterning, anomaly, and re-vision (in the most literal sense). In both of the literary test cases documented in the papers in this session, text mining has produced compelling insights that already provide the basis for more traditional scholarly interventions—papers and articles—in their respective subject fields. The technical environments featured in the papers likewise have promise in their own right and stand ready to support text analysis (Tamarind), structured text visualization (Maryland's adaptation of the InfoVis Toolkit), and a newly designed visual environment in support of the kind of complex, aggregative operations endemic to data mining (the Clear Browser).

In "Undiscovered Public Knowledge," Kirschenbaum et al. report on their experiments mining for patterns of erotic language in the poetry and correspondence of Emily Dickinson. This paper also describes significant components of the complete nora architecture, including the end-user visualization toolkit. In "Distinguished Speakers," Ramsay and Steger explore keyword extraction methods as a way of prompting critical insight. Using the particular case of Virginia Woolf's novel The Waves, they explore the use of the tf-idf formula and its variations for finding the "distinctive vocabulary" of individual characters in a novel. They also discuss their use of Tamarind (an XML preprocessor for scholarly text analysis used by the nora project) to make such investigations faster and easier. In "The Clear Browser," Ruecker, Rossello and Lord describe their attempt to create a visual interface design that is effectively positioned to be attractive for humanists. The goal of this sub-project is to help make the system accessible and interesting for scholars who might have an interest in the results of data mining, but are not immersed in the technology.

All authors listed in the papers have communicated their willingness to participate.

References

S. Downie, J. Unsworth, B. Yu, D. Tcheng, G. Rockwell and S. Ramsay (2005). "*A revolutionary approach to humanities computing*?: Tools development and the D2K data-mining framework." ACH/ALLC 2005.

"Undiscovered Public Knowledge": Mining for Patterns of Erotic Language in Emily Dickinson's Correspondence with Susan Huntington (Gilbert) Dickinson

Catherine PLAISANT
Human Computer Interaction Lab
Martha Nell SMITH
English and MITH, University of Maryland

Loretta AUVIL
NCSA, University of Illinois
James ROSE
Computer Science, University of Maryland
Bei YU
GSLIS, University of Illinois
Tanya CLEMENT
English, University of Maryland

This paper develops a rationale for "provocational" text mining in literary interpretation; discusses a specific application of the text mining techniques to a corpus of some 200 XML-encoded documents; analyzes the results from the vantage point of a literary scholar with subject expertise; and finally introduces a tool that lets non-specialist users rank a sample set, submit it to a data mining engine, view the results of the classification task, and visualize the interactions of associated metadata using scatterplots and other standard representations.

Text mining, or machine learning as it is also known, is a rapidly expanding field. Canonical applications are classification and clustering (Weiss 2005, Widdows 2004, Witten 2000). These applications are becoming common in industry, as well as defense and law enforcement. They are also increasingly used in the sciences and social sciences, where researchers frequently have very large volumes of data. The humanities, however, are still only just beginning to explore the use of such tools. In the context of the Nora Project, a multidisciplinary team is collaborating to develop an architecture for non-specialists to employ text mining on some 5 GB of 18th and 19th century British and American literature. Just as importantly, however, we are actively working to discover what unique potential these tools might have for the humanist.

While there are undoubtedly opportunities for all of the normative text mining applications in large humanities repositories and digital library collections, their straightforward implementation is not our primary objective with Nora. As Jerome McGann and others have argued, computational methods, in order to make significant inroads into traditional humanities research, must concern themselves directly with matters of interpretation (2001). Our guiding assumption, therefore, has been that our work should be provocational in spirit—rather than vocational, or merely utilitarian—and that the intervention and engagement of a human subject expert is not just a necessary concession to the limits of machine learning but instead an integral part of the interpretative loop. In important respects we see this work as an extension of insights about modeling (McCarty 2004), deformation (McGann 2001), aesthetic provocation (Drucker 2004), and failure (Unsworth 1997). It also comports with some of the earliest applications of data mining, such as when Don Swanson associated magnesium deficiency with migraine headaches, an insight provoked by patterns uncovered by data mining but only subsequently confirmed through a great deal of more traditional medical testing (Heast 1999).

We began with a corpus of about 200 XML-encoded letters comprising correspondence between the poet Emily Dickinson and Susan Huntington (Gilbert) Dickinson, her sister-in-law (married to her brother William Austin). Because debates about what counts as and constitutes the erotic in Dickinson have been primary to study of her work for the last half century, we chose to explore patterns of erotic language in this collection. In a first step our domain expert classified by hand all the documents into two categories "hot" and "not hot." This was done in order to have a baseline for evaluation of the automatic classifications to be performed later.

We then developed an exploratory prototype tool to allow users to explore automatic classification of documents based on a training set of documents classified manually. The prototype allows users to read a letter and classify it as "hot" or "not-hot" (Fig 1). After manually classifying a representative set of examples (e.g. 15 hot and 15 not-hot documents) this training set is submitted to the data mining classifier. For every other letter in the corpus, users can then see the proposed classification, review the document, and accept or change the proposed classification. The words identified by the data mining as possible indicators of erotic language are highlighted in the text of the document.

Importantly, this process can be performed in an iterative fashion as users improve the training set progressively and re-submit the automatic classification. Currently results are presented in the form of a scatterplot which allows users to see if there is any correlation between the classification and any other metadata attribute of the letters (e.g. date, location, presence of mutilation on the physical document, etc.) Users can see which documents have been classified by hand (they are marked with triangles) and which have been categorized automatically (they appear as a circle). Letters that have been classified as not-hot always appear in black, and in color for hot,

making it easy to rapidly spot the letters of interest.

A key aspect of our work has been to test the feasibility of this fairly complex distributed process. The Web user interface for manual and automatic classification is a Java Web Start application developed at the University of Maryland, based on the InfoVis Toolkit by Jean-Daniel Fekete (2004). It can be launched from a normal Web page and runs on the user's computer. The automatic classification is performed using a standard Bayesian algorithm executed by a data mining tool called D2K, hosted at the University of Illinois National Center for Supercomputing Applications. A set of web services perform the communication functions between the Java Interface and D2K. The data mining is performed by accessing a Tamarind data store provided by the University of Georgia, which has preprocessed and tokenized the original XML documents. The entire system is now functional.

What of the results? The textual critic Harold Love has observed of "undiscovered public knowledge" (consciously employing the aforementioned Don Swanson's phrase) that too often knowledge, or its elements, lies (all puns intended) like scattered pieces of a puzzle but remains unknown because its logically related parts are diffused, relationships and correlations suppressed (1993). The word "mine" as a new indicator identified by D2K is exemplary in this regard. Besides possessiveness, "mine" connotes delving deep, plumbing, penetrating--all things we associate with the erotic at one point or another. So "mine" should have already been identified as a "likely hot" word, but has not been, oddly enough, in the extensive critical literature on Dickinson's desires. "Vinnie" (Dickinson's sister Lavinia) was also labeled by the data mining classifier as one of the top five "hot" words. At first, this word appeared to be a mistake, a choice based on proximity to words that are actually erotic. Many of Dickinson's effusive expressions to Susan were penned in her early years (written when a twenty-something) when her letters were long, clearly prose, and full of the daily details of life in the Dickinson household. While extensive writing has been done on the blending of the erotic with the domestic, of the familial with the erotic, and so forth, the determination that "Vinnie" in and of itself was just as erotic as words like "mine" or "write" was illuminating. The result was a reminder of *how* or *why* some words are considered erotic: by their relationship to other words. While a scholar may un-self-consciously divide epistolary subjects within the same letter, sometimes within a sentence or two of one another, into completely separate categories, the data mining classifier will not. Remembering Dickinson's "A pen has so many inflections and a voice but one," the data mining has made us, in the words of our subject expert, "plumb much more deeply into little four and five letter words, the function of which I thought I was already sure, and has also enabled me to expand and deepen some critical connections I've been making for the last 20 years."

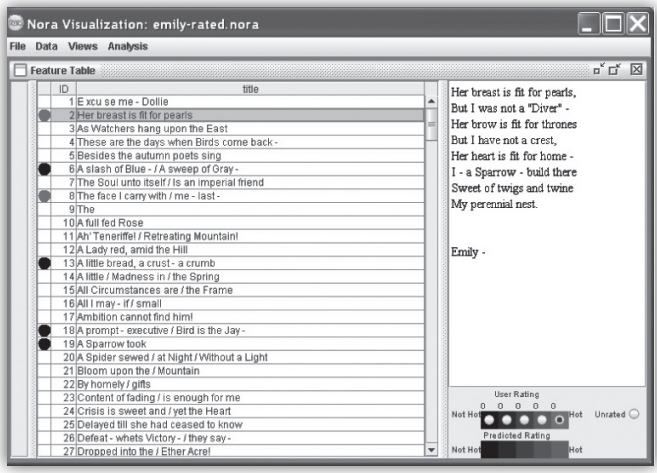


Figure 1: Users can select a document in the collection (here "Her breast is fit for pearls") and read the document. They can then classify it as hot (red) or not-hot (black), which helps build the training set.

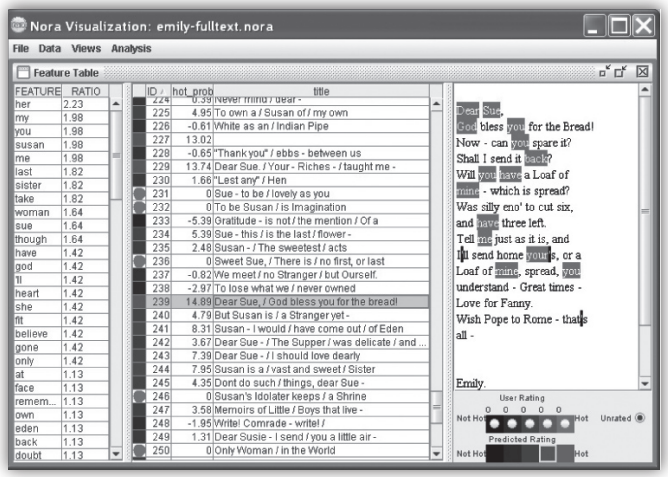


Figure 2: After requesting that the remaining documents be automatically classified, purple color squares are placed next to each document that had not been classified manually. Bright colors mean that the data mining suggests that the document might be "hot" and black means "not hot". On the most left pane, a list of words is provided, with the words found to be more representative of the hot documents of the training set listed at the top.

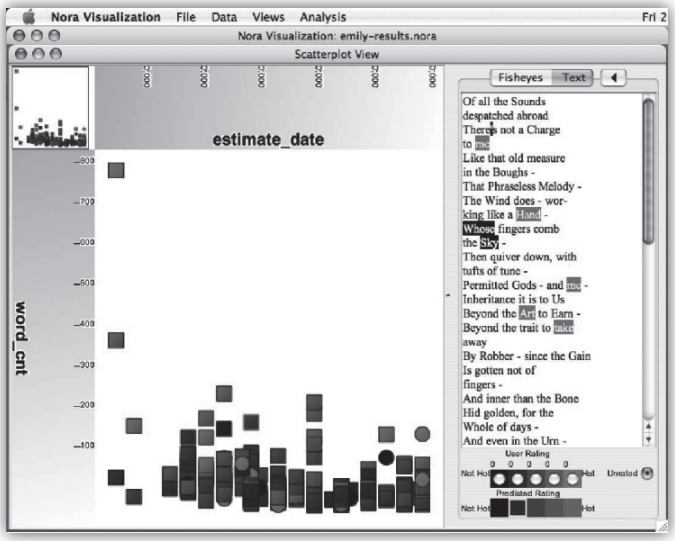


Figure 3: This is an alternate "scatterplot" view of the collection. Each dot represent a document. Time (i.e. the median of the estimated date range) is mapped on the X axis, and the length of the document is mapped on the Y axis. Color represents hotness. The same color coding is used. We can see that the longer documents were written earlier on. The display also suggests that there is no correlation between time and hotness, and no particular time periods where significantly more hot documents were written. Zooming is possible to inspect particular clusters.

References

Drucker, J. and B. Nowviskie. (2004). "Speculative Computing: Aesthetic Provocations in Humanities Computing." In S. Schreibman, R. Siemens, and J. Unsworth (eds.), *The Blackwell Companion to Digital Humanities* (pp. 431-447). Oxford: Blackwell Publishing Ltd.

Fekete, J-D. (2004). "The Infovis Toolkit." In *Proceedings of the 10th IEEE Symposium on Information Visualization* (pp. 167-174). Washington DC: IEEE Press.

Hearst, M. (1999). "Untangling Text Data Mining." At <http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html>.

Love, H. (1993). *Scribal Publication in Seventeenth-Century England*. Oxford: Clarendon Press.

McCarty, W. (2004). "Modeling: A Study in Words and Meanings." In S. Schreibman, R. Siemens, and J. Unsworth (eds.), *The Blackwell Companion to Digital Humanities* (pp. 254-270). Oxford: Blackwell Publishing Ltd.

McGann, J. (2001). *Radiant Textuality: Literature After the World Wide Web*. New York: Palgrave.

Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M.G, Smith, M.N, Clement, T. and Lord, G. (2006). *Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces*, to appear in the *Proceedings of the Joint Conference on Digital Libraries* (JCDL 06).

Unsworth, J. (1997). "The Importance of Failure." *The Journal of Electronic Publishing* 3.2. At < http://www.press.umich.edu/jep/03-02/unsworth.html>.

Weiss, S., et al. (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer.

Widdows, D. (2004). *Geometry and Meaning*. Stanford: CLSI Publications.

Witten, I. and E. Frank. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Diego: Academic Press.

Distinguished Speakers: Keyword Extraction and Critical Analysis with Virginia Woolf's *The Waves*

Stephen RAMSAY

Sara STEGER

Department of English, University of Georgia

From the earliest epistolary novels of the eighteenth century to the stream-of-consciousness narratives of the twentieth, English novelists have constructed narratives in which a single story is told from a variety of different first-person viewpoints. The motivations

for this technique are as ramified as the variations of the technique itself; some authors have used it to demonstrate the contingent nature of subjectivity, while others have employed the technique merely as a way of increasing dramatic irony and tension. In most cases, the individuated chorus of speakers is distinguished stylistically. One might compare, in this context, the scientific formality of the male characters in *Dracula* with the more personable journalistic endeavors of Mina and the twittering effusions of Lucy; the unprepossessing nobility of Hartright with the urbane wickedness of the Count in *The Woman in White*; the fractured, acausal narrative of Benjy Compson with the neurotic eloquence of his brother Quentin in *The Sound and the Fury*.

Virginia Woolf's 1931 novel *The Waves* employs this approach to narrative in order to trace the lives of six friends from early childhood to old age. The characters each tell their stories at seven distinct stages of their lives. Each monologue is clearly delineated, and all six characters share some of the same experiences at different points in the narrative.

Yet some critics claim these characters are not differentiated from one another stylistically, and that the distinguishing features of the characters have more to do with the complex symbolic landscape each one inhabits -- a motif foregrounded by the lack of stylistic differentiation. J. Guiget, for example, maintains that "these are not voices, in the sense that they are not differentiated. But for the 'Bernard said' or 'Jinny said' that introduces them, they would be indistinguishable; they have the same texture, the same substance, the same tone" (283). Likewise M. Rosenthal suggests that although the speeches are highly stylized, the language is undifferentiated. He goes on to say that "although the different speakers all have different points of view and preoccupations, they use the same kind of sentence rhythms and employ similar kinds of image patterns" (144). For these critics, the six characters are not characters at all, but voices indistinguishable by means of language or imagery.

There is, however, an opposing critical position that stresses the importance of stylistic differentiation among the characters in the novel (the specific contours of this difference being the main point of scholarly contention). Charlotte Mendez draws the line of differentiation along the gender axis. Hermione Lee emphasizes the distinction between the worldly speakers and the more solitary speakers. Susan Gorsky emphasizes the individuality of the characters through a clustering of their primary images. According to her research, "the speech of each character is made distinct within the mask of the formal monologue by the repetition of key phrases and images. Diction varies from one speaker to the next because of the words repeated in the image patterns" (454).

With this critical backdrop, certain questions naturally emerge. Do the characters employ similar image patterns or distinctive language patterns? Is there a way to group characters based on similarities in their speeches? Are there six voices in the novel or is there only one?

Our goal was not to adjudicate the matter, but to seek further entry points into these axes of intelligibility. Knowing that vocabulary (symbolic or not) would be one vector of interest, we employed several variations of the tf-idf formula as a way to separate the characters. Tf-idf -- a popular formula in information retrieval -- weighs term frequency in a document against the frequency of that term throughout a corpus. By assigning weights to terms, it attempts to re-fit a word frequency list so that terms are not distributed according to Zipf's law. We compared every word token in each of the six characters to every other character's vocabulary, and used the resulting lists of "distinctive terms" as the basis for further reflection on the individuation of character in Woolf's novel.

In generating our results, we had recourse to Tamarind (one of the software subsystems for the nora project). This system, which acts as an XML pre-processor for scholarly text analysis, tokenizes, parses, and determines part-of-speech markers for each distinct token in a corpus. Using this system as a base, we were able to conduct comprehensive term-comparisons using only 50 lines of code (in Common Lisp). We therefore think of this project as a test case for the feasibility of using Tamarind as way to simplify complex text analysis procedures of the sort envisioned by the larger nora project.

In this paper, we present the results of our investigation into Woolf's narrative, while also looking at the ways in which the software architecture for nora enabled us to undertake the study quickly and easily. Drawing on similar work with tf-idf in digital humanities (e.g. Rydberg-Cox's work with Ancient Greek literature), we suggest some of the ways in which the results of keyword

extraction algorithms might be further processed and visualized. Finally, and perhaps most importantly, we discuss the ways in which the computerized generation of "suggestive pattern" can enable critical reflection in literary study.

References

Gorsky, Susan (1972). "The Central Shadow: Characterization in *The Waves*." *Modern Fiction Studies* 18.3: 449-466.

Guiguet, Jean (1965). *Virginia Woolf and Her Works*. London: Hogarth.

Lee, Hermione. Virginia Woolf. New York: Vintage, 1996.

Mendez, Charlotte Walker (1994). "Creative Breakthrough: Sequence and the Blade of Consciousness in Virgina Woolf's The Waves." *Virginia Woolf, Critical Assessments*. Ed. Eleanor McNees. Mountfield, England: Helm Information.

Rosenthal, Michael (1979). *Virginia Woolf*. London: Routledge.

Rydberg-Cox, Jeffrey A. (2002). "Keyword Extraction from Ancient Greek Literary Texts ." *LLC* 17: 231-44.

Woolf, Virginia (1931). *The Waves*. New York: Harcourt.

The Clear Browser: Visually Positioning an Interface for Data Mining by Humanities Scholars

Stan RUECKER
Humanities Computing in English and Film Studies, University

Ximena ROSSELLO
Dept of Art and Design, University of Alberta

Greg LORD
Maryland Institute for Technology in the Humanities (MITH), University of Maryland

Milena RADZIKOWSKA
Centre for Communication Studies, Mount Royal College

We describe in this paper a strategy for interface design based on the concept of visual positioning. We apply this strategy to the design of an interface for the Nora project, which presents a unique opportunity to create tools to accommodate a powerful technology-data mining-to a new group of users-humanities scholars.

The goal of the Nora project is to apply state-of-the-art data mining processes to a wide range of problems in the humanities (Unsworth 2005), not only in the service of hypothesis testing, but also as a means of contributing to hypothesis formulation (Shneiderman 2001; Ramsay 2003). In both of these cases, however, the question arises of how to make the power of data mining for text collections accessible to academics who are neither mathematicians nor computer programmers. Typical interfaces for data mining operations involve either command lines, such as are used in working in UNIX, or else GUIs, the visual positioning of which frequently places them in a technical domain-many resembling the interfaces used in software compilers. For humanities scholars, it is necessary to consider alternative designs that attempt to adopt a visual position that is at once more congenial and more appropriate for humanists, while at the same time sacrificing as little as possible of the functional control of the underlying system.

The concept of visual positioning has become widespread in the visual communication design community. An early formulation of the principle was provided by Frascara (1997) who pointed out that since one of the primary goals of the graphic designer is to improve communication, it is necessary to consider the visual environment and visual preferences of the users in order to increase the success of the design in communicating with them. The application of this concept to interface design suggests that there are going to be designs that are more or less successful for a particular group of users, and that the same designs won't necessarily be successful to the same degree with a different group that does not share the same visual position.

In connection with the Nora project, the necessary communication is between the technical mechanism

of the data mining processes and the potential user-the humanities scholar. A typical data mining operation consists of the following stages:

1) the system provides the user (in this case, a scholar) with a sample of documents from the collection

2) the scholar chooses among the sample documents those which are of interest for a particular study. In the two Nora project examples, a sample of poems from a collection of Emily Dickinson was rated in terms of erotic content, and a sample of novel chapters was rated according to their instantiation of the concept "sentimentalism."

3) the system performs a set of "feature extraction" actions in order to determine shared characteristics of the selected documents

4) the scholar examines the shared characteristics and iteratively adjusts the result as necessary

5) the system applies the resolved characteristics to the larger collection in order to automatically identify similar documents

6) the scholar studies both the shared characteristics and the result set, often by using a visualization tool (in Nora, the InfoVis toolkit).

We call the interface intended to facilitate this process the clear browser. It is based on the idea of rich-prospect browsing, where some meaningful representation of every item in the collection is combined with a set of tools for manipulating the display (Ruecker 2003). In this case, the primary tools are in the form of a set of "kernels" which encapsulate in visual form the results of the data training stage. The kernels allow a simple means of storing the results of feature extraction processes for further modification or use, and also give the user a simple mechanism for applying the process, by dragging and dropping the kernel within the representation of all the collection items (Figure 1). The effects of the kernel are to visually subset the collection items into two groups-selected and unselected-so that the user can subsequently access the items in the selected subset. The design also allows for combinations of kernels, and for a single kernel to provide multiple functions, including not only subsetting the items, but also adding further grouping or sorting functions, as well as changes to the form of representation.

Figure 1. The Clear Browser provides a number of blank kernels that can be configured by the user through a data mining "training" process. These kernels can then be applied to the larger collection by dragging and dropping them. This sketch shows a total collection of 5000 author names, with a subset selected by the kernel.

One of the important aspects of the visual positioning for humanities scholars is the proposed form of the meaningful representation of the individual items in the collection. These items are each a piece of text, and together they form a large body of text that is displayed on screen as the default interface. It perhaps goes without saying that humanities scholars are comfortable with text, whether in print or on screens, and the choice to represent collection items with text can therefore contribute to their ability to interpret quickly and intuitively what is happening with a system that might otherwise be unfamiliar or disorienting.

For purposes of illustration, it might be helpful at this point to introduce a scenario involving changes to the form of representation. Such a change might be introduced by the system in connection with a sorting action. For example, if the items in the collection are initially represented as the titles of poems, and the user elects to sort the selected poems by date of first publication, it would typically be useful at that point to add the date to the name of each poem. This addition would constitute a change to the individual representations of items. Alternatively, in cases where the user prefers to group the items rather than sort them,

the additional information might be attached to the entire group in the form of a group label, in which case the representations of the individual items in the group would remain unchanged.

Another aspect of the visual positioning is the animated actions of the kernels, which interact with the field of representations with an effect like oil and water. The animation of the movement of the text items, which move to the periphery of the display or the centre of the area associated with the kernel, provides two kinds of cognitive reassurance. First, the user has a sense of being able to follow the action of the data mining process as encapsulated in the kernel. Second, the animated transitions of the text items provide reassurance that the system is rearranging the collection without adding or subtracting any items. This second factor is particularly important in cases where one of the other functions of the kernel is to add or subtract components from the meaningful representation. By animating the movements and changes in discrete steps, the interface helps make the results of the process understandable. The animated actions of the items become part of the visual positioning, not because cognitive reassurance isn't important for all users, but because some users can benefit more than others from having it provided in this form.

References

Frascara, Jorge (1997). *User-centered Graphic Design*. London: Taylor and Francis.

Ramsay, Stephen (2003). "Toward an Algorithmic Criticism." *Literary and Linguistic Computing*. 18.2.

Ruecker, Stan (2003). *Affordances of Prospect for Academic Users of Interpretively-tagged Text Collections*. Unpublished Ph.D. Dissertation. Edmonton: University of Alberta.

Shneiderman, Ben (2001). "Inventing Discovery Tools: Combining Information Visualization with Data Mining." Keynote for Discovery Science 2001 Conference, November 25-28, 2001, Washington, DC.

Unsworth, John (2004). "Forms of Attention: Digital Humanities Beyond Representation." Paper delivered at CaSTA 2004: The Face of Text. 3rd conference of the Canadian Symposium on Text Analysis, McMaster University, Hamilton, Ontario. November 19-21 2004.

COMPARING AGGREGATE SYNTAXES

John NERBONNE
University of Groningen, The Netherlands

Franz MANNI
Musée de l'Homme, Paris

Recently, large and representative databases have become available which record how languages and dialects differ syntactically, i.e. with respect to the way in which words and phrases combine. Barbiers, Cornips and van der Kleij (2002) and Longobardi are examples of such data collections, and several more are in construction, and should be available presently. This availability of large amounts of digitized, controlled syntactic data enables several new questions to be addressed, including how syntactic features are geographically distributed (dialectology and geolinguistics, see Spruit, 2005), to what degree syntactic features associate with one another (typology), and the degree to which syntactic features follow phonological and lexical features in their geographic or linguistic (typological) patterning.

Perhaps most intriguingly both Guardino & Longobardi (2005) and Dunn et al. (2005) postulate that syntactic features are more resistant to change than other linguistic properties, so that careful examination of shared syntactic features and especially shared syntactic innovation should be not only a welcome addition to the techniques of historical reconstruction in linguistics, which rely primarily on phonetic and morphological evidence, but potentially an improvement.

None of these investigations is conceivable without extensive computational support. Not only is the data digitized, but the analytical procedures, the statistical analysis of the results and their visualization all require significant processing time.

The purpose of this special session will be to present representative studies exploring the various ways in which these syntactic databases are being exploited, including ways which might engage other students of the history of human culture.

There will be three papers, one each by the groups at

Nijmegen, Trieste and Groningen-Amsterdam. Michael Dunn of Nijmegen will present further aspects of the Nijmegen group's work, featured earlier this year in Science, aimed at using syntactic comparison to reconstruction language history in New Guinea, Giuseppe Longobardi will present new aspects of his work (with Gianollo and Guaridano) exploring especially the value of abstract syntactic features in historical reconstruction, and Marco Spruit and John Nerbonne will present their work (with Heeringa) comparing syntactic distances on the one hand with lexical and phonetic differences on the other.

References

Barbiers, S., L. Cornips & S. van der Kleij (eds). (2002) Syntactic Microvariation. Electronic publication of Meertens Institute and NIWI. URL: http://www.meertens.knaw.nl/books/synmic

Dunn A.M., Terrill A., Reesink G., Foley R. & Levinson S.C. (2005) Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science*, (309), 2072-2075.

Gianollo, C., C.Guardino, & G.Longobardi (2004, in press) Historical Implications of a Formal Theory of Syntactic Variation. In S.Anderson & D.Jonas (eds.) *Proc. of DIGS VIII*.

Guardino, C. (2005) Parametric Comparison and Language Taxonomy. In M.Batallori, M.L.Hernanz, C.Picallo, and F.Roca (eds.) *Grammaticalization and Parametric Variation*. Oxford: OUP. pp.149-174.

Spruit, M. (2005) Classifying Dutch dialects using a syntactic measure: The perceptual Daan and Blok dialect map revisited. In J. Doetjes & J. van de Weijer (eds.*) Linguistics in the Netherlands*. Benjamins, Amsterdam. 179-190.

Contact and Phylogeny in Island Melanesia

Michael DUNN, Ger REESINK
Max Planck Institute for Psycholinguistics, Nijmegen

Due to the continual process of erosion of the lexical material in a language,the phylogenetic signal detectable by the comparative method inevitably disappears into noise at some point in the past. Dunn et al. (2005) presents some results of a project investigating deep time relationships between the non-Austronesian group of languages (the so-called 'Papuan' languages) in Island Melanesia, a group of languages which (i) could plausibly be related (especially when considering recurrent typological similarities across most of the group), but which (ii) cannot be shown to be related using standard linguistic methods. Methods from computational biology have been adapted to address questions of language history, treating structural features of language as genealogically transmitted traits.

Island Melanesia is a fascinating natural laboratory for the study of language change. From 30000 to 3000 years before present these islands (the Bismarck Archipelago, Bougainville, and the Central Solomons) formed the furthermost limit of human dispersal into the south Pacific. In the last 3000 years the Austronesian expansion encompassed near Island Melanesia, and continued eastwards into the Pacific. Currently 90% of the languages in near Island Melanesia are members of the Oceanic subgroup of the Austronesian family. With a few interesting exceptions, these languages fall into reconstructable genealogical relationships using standard comparative methods. The remaining 10% of the languages of the region are the hitherto unrelatable Papuan remnants of the pre-Austronesian linguistic diversity. The Papuan languages are largely out of contact with each other.

The stability of individual structural features relative to the lexicon is known to be variable, yet it takes rather special conditions of prolonged contact for extensive exchange of structure to take place. While we do not have evidence to support a claim that language structure is universally more conservative than lexical form based relationships, it is plausible that in some cases it would be so, and the languages of Island Melanesia seems likely candidates. This is in any case an empirical question.

The presence of the Oceanic Austronesian languages in the region gives us a control group of languages with known phylogeny from standard methods. Structural phylogenetic methods use an independent set of data to the lexical data of the comparative method, so once a computational tool can be shown to detect a phylogenetic

signal that matches the known signal of the control, it becomes scientifically interesting to apply the tool to the target group of lexically unrelatable Papuan languages. The set of relationships shown by the linguistic structural data are geographically plausible, and have provided hypotheses for testing in human genetics.

To date the main method we have used to investigate structural relationships between languages is maximum parsimony, which emphasises the phylogenetic component of the historical signal carried by structural traits. There is no doubt that that a process of reticulation through borrowing and contact must also explain some of the variation. Any set of linguistic structural data will show some reticulation (in linguistic terms, homology due to contact-induced change and chance convergence). In biological systems these phenomena are relatively rare, and the methods for investigating them are newer. The NeighborNet method implemented in the SplitsTree package (Bryant and Moulton, 2004) is emerging as the current standard (used e.g. by Bryant, Filimon and Gray 2005; Cysouw and coworkers with the WALS data at MPI EVA, Leipzig). These networks concur with the observations of descriptive linguists that Oceanic-Papuan ontact has had a significant influence on some groups of languages within Island Melanesia. Deeper exploration of the typological database used shows--not unexpectedly--that traits are differentially involved in horizontal transfer between genera, leading to the possibility of developingdata-driven methods for statistically compensating for the conctact signal and amplifying the signal of phylogeny.

While for a long period strict adherence to the comparative method has provided a needed rigor to counterbalance other more dramatic but ultimately unverifiable methods (e.g. the 'megalo-comparativists'; Matisoff 1990), advances in areal linguistics and statistical approaches now allow us to investigate linguistic change outside the scope of the comparative method.

References

Bryant, D. & Moulton, V. 2004. NeighborNet: an agglomerative algorithm for theconstruction of phylogenetic networks. *Molecular iology and Evolution* 21(2):255-265.

Bryant, D., Filimon, F. & Gray, R. 2005. Untangling our past: Languages, Trees, Splits and Networks. In: *The Evolution of Cultural Diversity: Phylogenetic Approaches*. ed. by R. Mace, C. Holden, S. Shennan. UCL Press.

Dunn M., Terrill A., Reesink G., Foley R. & Levinson S.C. 2005. Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science*, 309:2072-2075 .

Matisoff, J.A. 1990. On megalo-comparison: a discussion note. *Language* 66:106-20.

Is a 'History and Geography of Human Syntax' meaningful?

C. GIANOLLO, C. GUARDIANO, Guiseppe LONGOBARDI
University of Trieste

In addition to its theoretical impact, the development of molecular biology has brought about the possibility of extraordinary scientific progress in the historical study of classification and geographical distribution of different species and different human populations (cf. Cavalli Sforza et al., 1994).

We want to suggest that parametric theories of linguistic variation in generative grammar can prompt analogous progress in the study of the history and geographical distribution of different language families.

Thus, this work aims at unifying two traditionally unrelated lines of investigation:

- the formal study of syntactic variation in generative grammar
- the reconstruction of historical relations among languages (phylogenetic taxonomy)

The pursuit of this approach will be argued to seriously question the traditional belief in the orthogonality of grammatical typology and language diachrony/genealogy, enabling us to tentatively suggest a positive answer to the following problem, that we conventionally label

'Humboldt's problem':

> Are the typological and the genealogical (phylogenetic) classifications of languages significantly isomorphous?

From the second half of the 19th century on, three different types of enterprises attempted to classify languages and/or populations separated for centuries or millennia into historically significant families:

A) for relatively shallow time depths and in particularly favored cases, the classical linguistic comparative method, based on inspection of the lexicon, can often provide sharp taxonomic conclusions, immune from the need of serious probabilistic evaluation, since the relatedness hypotheses are warranted by few patently highly improbable phenomena, most notably recurrent (optimally 'regular') sound correspondences. These properties largely solve the problem of a safe choice of comparanda, allowing for solid conclusions.

B) Beyond the classical one, the only other linguistic method so far proposed is Joseph Greenberg's (e.g. 1987, 2000) mass comparison; still based on the lexicon, it suggests more far-reaching but much less rigorous and widely acceptable taxonomic conclusions, because the very choice of the compared entities, based on pretheoretical resemblance among arrays of words, is less safe from chance. Obvious probabilistic questions, often of unmanageable complexity, arise and receive controversial answers.

C) A third comparative practice stems from a different discipline, population genetics (cf. Cavalli Sforza et al., op. cit.): no question arises here about the comparability of the basic entities, since they are drawn from a finite and universal list of biological options: a blood group must be compared to the same blood group in another population, obviously, not to other sorts of genetic polymorphisms. The only issue concerns the statistical and empirical significance of the similarities discovered. This is why population genetics is considered so useful to complement linguistics in the task of classifying populations and languages.

We propose that the contribution of linguistics proper to such issues can be completely renovated on the grounds of parametric generative theories, which in principle allow one to bring modern cognitive science to bear on issues of cultural variation and historical explanation. Since parameters form a finite and universal list of discrete biological (though culturally set) options, they resemble the set of polymorphisms studied by population genetics and potentially enjoy similar (and perhaps even greater) formal advantages, overcoming in principle all questions on the choice of comparanda affecting linguistic methods based on the vocabulary. On the other side, the a priori probative value of parametric comparison is mathematically very high: e.g. just 30 binary independent parameters generate 230 languages = 1,073,741,824. The probability for two languages to coincide in the values of 30 independent parameters with binary equiprobable values = 1/230, of three languages = (1/230)2, i.e. less than one in one billion billions.

We will test and exploit such a potential by establishing exact comparisons of parameter values among some languages whose degree of cognation is independently known, in order to prove the effectiveness of the method to provide historically correct taxonomic insights before applying it to controversial cases.

For the past fifteen years, a number of scholars have studied the parametric variation of the structure of nominal phrases in several languages. Relying on this and new specific work and following the MGP method of Longobardi (2003: relatively many parameters in relatively many languages in a single limited subdomain), we have worked out a preliminary list of 50 binary parameters affecting DP-internal syntax and tested their values in over 20 ancient and contemporary varieties drawn from several Indoeuropean and non-Indoeuropean subfamilies (including Modern Italian, French, European Portuguese, Latin, Classical Greek, Modern Greek, Gothic, Old English, Modern English, German, Bulgarian, Serbo-Croat, Arabic, Hebrew, Hungarian, Wolof, among others). Each relevant parameter has been tentatively set for such languages, obtaining up to 50 precise correspondence sets of parameter values, and for every pair of languages we could arithmetically count identities and differences. In our formalism, the relative distance between any two languages is expressed by a coefficient, which consists of an ordered pair of positive integers <i,d>, where i is the number of identities and d the number of differences. This procedure of lexically blind comparison, coupled with especially designed empirical and statistical methods, will be argued to prove adequate to generate the essentially correct phylogenetic tree, as resulting from traditional methods

of lexical comparison. This is why the answer to Humboldt's problem above appears (perhaps surprisingly) tendentially positive.

The elaboration of a compact variation model of a whole syntactic subdomain, even though for the mere purposes of historical taxonomy, begins, in turn, to yield a number of insights for parameter theory itself from both a synchronic and a diachronic point of view:

- we will show how parameters, even in a limited area of grammar, appear to be tightly interrelated and not freely orthogonal to each other (cf. Fodor 2000): some parameters become irrelevant (in different senses, to be technically distinguished) in a language due to the setting of other parameters (or to variation in the composition of the lexicon: Kayne 2003);
- parameters can be evaluated according to their relative historical weight:

1) certain parameters may understate or overstate identities and differences because they trigger a high number of irrelevant settings in the other parameters. Their diachronic resetting will be much more consequential (catastrophic in Lightfoot's 1999 sense) than that of others;
2) some parameters appear to be taxonomically more significant than others because they are diachronically more stable. The possibility that this may be due to different degrees of exposure to external triggers and other modules at the interface, along the lines of Keenan's insight about Inertia, will be explored.

This line of investigation will be construed as a peculiar way to progress 'beyond explanatory adequacy' in generative grammar (Chomsky 2001), leading to a better understanding of both classical and new issues in the cultural and natural history of the language faculty.

References

Cavalli Sforza, L. P. Menozzi, and A. Piazza (1994). *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Chomsky, N. (2001). Beyond explanatory adequacy. Eliminating labels. *MIT Working Papers in Linguistics*.

Fodor, J. (2000) *The Mind doesn't Work that Way :The Scope and Limits of Computational Psychology*. Cambridge: MIT Press.

Greenberg, J. (1987). *Language in the Americas*. Stanford: Stanford University Press.

Greenberg, J. (2000). *Indo-European and Its Closest Relatives: The Eurasiatic Language Family. Volume I: Grammar*. Stanford: Stanford University Press.

Kayne, R. 2003. "Antisymmetry and Japanese", *English Linguistics* 20. 1-40.

Lightfoot, D. (1999) *The Development of Language: Acquisition, Change, and Evolution*. Oxford: Blackwell.

Longobardi, G. (2003) Methods in Parametric Linguistics and Cognitive History. *Linguistic Variation Yearbook* 3, pp. 103-140.

Associations among Linguistic Levels

Marco SPRUIT
Meertens Institute, Amsterdam

Wilbert HEERINGA, John NERBONNE
University of Groningen

We are now in a position to assess the dialectometric distances among fairly many sites at three different linguistic levels: pronunciation (phonetics and phonology), lexicon (or vocabulary), and syntax. We shall refer to pronunciation differences as phonological, even if they involve subphonological variation as well. We measure lexical and syntactic differences at a nominal level, effectively using techniques introduced by Seguy (1971) and Goebl (i.a. 1982), and we measure pronunciation differences numerically, using Levenshtein distance (Nerbonne, Heeringa and Kleiweg, 1999; Heeringa, 2004). The novelty of this paper consists first in the opportunity to include syntax among the linguistic levels we analyze, and second, in its attention to potential, mutually structuring elements among the linguistic

levels.

While most linguists would predict that vocabulary is more volatile than pronunciation and syntax, and might predict that lexical choice should show little association with other linguistic levels, there have been predictions linking pronunciation with syntactic properties (Donegan & Stampe, 1983). Both pronunciation and syntax are highly structured systems, within which a single linguistic parameter might lead to a multitude of concrete and measurable effects.

We address two research questions in the present paper, the first of which is fairly straightforward:

1) To what degree are aggregate phonological, lexical, and syntactical distances associated with one another when measured among varieties of a single language?

   1a) Are syntax and phonology more strongly associated with one another than either (taken separately) is associated with lexical distance?

To answer the questions above, it is sufficient to calculate correlation coefficients among the distance measurements for the three linguistic levels. This is a reasonable measure of the degree to which the three linguistic levels are associated.

It would be a mistake, however, to interpret any such correlation as influence without checking for the influence of a third factor. This is not merely the methodological reminder that one ought not interpret correlation as causation. More specifically, geography has independently been shown to that correlate highly with each of these linguistic levels, and it is quite plausible that geography could influence each of the levels separately, leading to the impression of structural influence between them. We suggest that this should be regarded as a null hypothesis, i.e. that there is no influence among the various linguistic levels. the This leads to the second research question we wish to address in this paper:

2) Is there evidence for influence among the linguistic levels, even once we control for the effect of geography?

   2a) Do syntax and phonology more strongly influence one another than either (taken separately) influences or is influenced by lexical distance?

We attack these latter questions in multiple regression designs, checking for the effects of linguistic levels on one another once geography is included as an independent variable.

References

Donegan, P., and D. Stampe (1983). Rhythm and the holistic organization of language structure. In: Richardson, J. F. et al. (eds.) *Papers from the Parasessioon on the interplay of phonology, morphology and syntax*. Chicago: Chicago Linguistic Society, 337--353.

Goebl, H. (1982). *Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*, Wien: Austrian Academy of Science.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. U.Groningen, Ph.D. Thesis.

Nerbonne, J., W. Heeringa, and P. Kleiweg (1999). Edit Distance and Dialect Proximity. In: David Sankoff and Joseph Kruskal (eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Stanford: CSLI Press, pp.v-xv.

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. In: *Revue de linguistique romane*, 35: 335--357.

DIGITAL RESEARCH OR DIGITAL ARTS?

Marcel O'GORMAN

Digital Media Studies, University of Detroit

To date, humanities computing has been entrenched in an archival sensibility. While the technologies involved in this field may be innovative, they are still geared toward the production of traditional research papers rooted in the materiality, methods and ideology of print culture. I suggest that humanities computing should not only focus on the preservation of the human record, but should also foster the invention of research methods more suitable to a digital culture. In addition, digital humanities projects should serve not only to better our understanding of human artifacts, but should also consider the impact of technology on the humanities and on human being itself. I will pursue this argument by presenting some of my own research projects, which blur the boundaries between the arts, critical theory, and the humanities. The projects in question may be viewed at the following URL's: http://www.spleenhouse.net; http://www.dreadmill.net .

For the time being I gave up writing--there is already too much truth in the world--an overproduction of which apparently cannot be consumed!

-Otto Rank, Letter to a Friend, 1933

In his brilliantly digressive response to the question, "Is Humanities Computing an Academic Discipline?" Geoffrey Rockwell quotes *Phaedrus* (http://www.iath.virginia.edu/hcs/rockwell.html). Within this famous passage, a Socratic account of the invention of writing, is King Thamus's response to the enthusiastic inventor, Theus:

> *O man full of arts, to one it is given to create the things of art, and to another to judge what measure of harm and of profit they have for those that shall employ them.*

My concern is that humanities computing, challenged by the need to financially sustain and legitimate itself in a rampant technocracy, has lost sight of the latter assignment noted by Theus. How do we measure the harm and profit that digital tools have for humanists, and for humanity in a more global sense? I do not expect to answer this question adequately in a single presentation, but I believe it is worthy of discussion--and in effect its absence would be conspicuous--at a conference in the digital humanities. I will approach the question by examining some specific projects in digital criticism that blur the boundaries between artistic practice and humanistic scholarship. In addition, I will argue that such projects point the way toward a new digital aesthetics in humanities research.

Following John Guillory, I argue in *E-Crit: Digital Media, Critical Theory, and the Humanities* (U of Toronto Press, 2006) that critical theory lent the humanities a much-needed aura of "rigor" at a time when academic institutions sought a fuller integration within the dominant technobureaucracy. I also argue that media technologies are serving the very same legitimating purpose for the humanities today, now that theory's aura has all but dissipated. Humanities computing, in this sense, is the new critical theory. But my hope is that the production of digital tools for the humanities is not simply an effect of what Heidegger called technology's "challenging forth." My hope as well, is that humanities computing, with its primary concern for preserving artifacts, will not fashion itself as a neo-traditional (though digital) backlash to the perceived threat imposed by critical theory on the canon. Finally, my hope is that computing humanists will draw on the methodological innovations, ontological suspicions, and phenomenological conceptions of technology, as put forth by Heidegger, Foucault, Derrida, etc., to develop research projects that both critique and make full use of the expressive forms available in digital media. Who will invent the scholarly methods suitable to a digital culture? The inventors of such tools should not only be given to create things of art, but should also be given to measure the harm and profit of such things.

As a former student of Gregory Ulmer, I came to *Phaedrus*--the first discourse on method--with a very clear purpose in mind: to understand how scholarly method is invented, and how it works within the communications apparatus that makes it materially possible. To date,

humanities computing projects, rooted in an "archival" sensibility, have done little more than increase the speed and efficiency of hermeneutic research methods developed in and for a print culture. My concern for the future of the humanities, then, is summarized in the following question, which I extend to the creators of digital humanities research tools, myself included: "**What will we do with all this information once it has been digitally archived?**" I believe that humanities computing can sustain and legitimate itself not only through the preservation of the human record, but also through the invention of new, digitally mediated research methods that serve to better our understanding not only of human artifacts, but of the impact of technology on the humanities and on human being. To that end, I will introduce two digital media projects, a performance and an installation, that attempt to combine humanistic scholarship, digital art, and the critique of technology.

Project 1: Dreadmill (http://www.dreadmill.net)

The "dreadmill" is a treadmill hardwired to a computer so that a runner's speed and heart rate control a multimedia show. In performances of 5-7 kilometres, I run on the dreadmill and discuss the "collusion of death and technology," drawing primarily on Heideggerian phenomenology, existentialism, and cultural theory. This performance, which I have given at several universities and art galleries in Canada and the United States, is designed to challenge preconceived notions of what defines humanistic scholarship and the dissemination of research. Not only does Dreadmill critique the immobility of the human body in a screen obsessed culture, it also critiques the sedentary practices of scholars, who submit themselves physically to the print apparatus. As a provocative contrast to the conventional lecture or the reading of a conference paper, Dreadmill radically implicates my body in the act of information delivery. I will show brief video footage of the performance and discuss its relevance to humanistic research practices.

Project 2: Spleenhouse (http://www.spleenhouse.net)

Spleenhouse, an extensive installation project, is designed to rescue the land, language, and practices of one of the few remaining French farmsteads in Ontario, Canada. The project involves relocating a greenhouse from a heritage site in LaSalle, Ontario, and relocating it in the nearby Art Gallery of Windsor, Ontario. Digital documentary footage of the few remaining Francophone farmers in the region will be projected onto the roof of the structure, providing light for the growth of vegetables inside the greenhouse. In the attempt to fuse form and content, this installation, provides a visceral environment for the encounter of rustic agrarian practices and digital techniques, rural simplicity and urban sophistication. I will show slides of the project, play a brief sample of the video footage, and discuss how Spleenhouse presents a new approach for disseminating humanities research in a digital culture. This project is especially suitable to this conference since the final destination of Spleenhouse is in Paris, France, and conference attendees will be able to visit it if they are interested.

For the past decade, I have incessantly been asked to define my work as either humanities research or digital art. This confusion regarding my area of specialization is something that I have consciously fostered. As I will argue throughout this presentation, humanities scholars have a great deal to gain by drawing on the practices and techniques of digital artists. The object of research in the projects outlined above might very well have been approached from a uniquely print oriented perspective, resulting in the production of academic essays and books—indeed, these projects have resulted in scholarly publications. But as I will discuss in my presentation, the media chosen to disseminate this research enhances it in several ways, providing an argument for the integration of artistic practices and materials in humanistic scholarship. In addition, the use of digital media in these projects is essential for engaging in a self-reflexive assessment of the impact of technology on the human condition.

Interactive Matter in the Arts and Humanities

Geoffrey ROCKWELL
Communication Studies and Multimedia, McMaster University

Digital Arts and Digital Humanities at first glance seem different, but do we know how they are different? One way to explore the intersections of the digital arts and humanities is to identify the location of differences and similarity, especially in research/creation

practices, dissemination, and projects.

This paper will tackle three issues around the differences between digital arts and humanities:

1. Practices or Methods. The practices of digital artists and digital humanists are evolving towards a common craft of interactive matter. The project cycles, funding issues and types of tools developed for interpretative purposes and aesthetic purposes have more in common than the practices of digital humanists have with their traditional colleagues. It is common in both the digital arts and humanities to involve students in apprentice-like opportunities on projects and train them that way.

Further, there are emerging similarities in the types of computing processes used in web projects in the arts and humanities. In this paper there will be demonstration of examples of digital art works and text visualization works to show the similarity and overlap. This overlap raises questions about the differences between an interpretative relationship with matter and a aesthetic relationship. Once we consider digital data as a form of material (McCullough, 2003) that is formed by the artist/craftsperson or by the humanist interpreter, we discover that the formative practices are similar even if the aims are not. It is instructive to imagine a topology of formative practices that might be common.

2. Research and Creation Dissemination. An arguable difference between the arts and humanities is the context of dissemination. The research of the humanities is typically disseminated through journals and monographs as part of a larger textual dialogue. Even if that which is studied is not itself textual, the community of researchers exchange information and disseminate their research through writings (and to a lesser extent through presented "papers".) By contrast the research/ creation of artists is disseminated through exhibits in galleries or museums, and it is typically disseminated in non-linguistic (or textual) form. If we ignore, for a moment, the textual ephemera of art dissemination (artist's statements, biographies, and curatorial documents), and also ignore all the textuality of art, the context and form of dissemination is different.

But will it always be so? The web as venue for dissemination is being used by both humanists and artists. The interactive web site as a form of dissemination is common to both digital artists and computing humanists. In certain interesting cases, like "TextArc" (www.textarc.org) it is not clear if the interactive site is art or academic work. For example, Matthew Mirapaul, Arts Columnist of the _New York Times_ is quoted on the "TextArc" site as writing, "TextArc evolves from an academic tool into a full-fledged work of digital art. ... This is the reading process made visible." (www.textarc.org/Reactions.html)

It is easy to say that such examples are the exceptions that make the rule of difference between academic interpretation and digital art. Given the importance of interface design to the field of computing, especially networked computing, it is harder to maintain this distinction in humanities computing projects. All digital humanities projects that aim to develop interactive works have to engage issues of design, though one could still argue that there is a difference between design and art that is based on a relationship between aesthetics and function. This paper will argue that there is a convergence through issues of usability and interface design between interactive art and humanities and that the site of this convergence is a common practical concept of interactive matter which has both form and content. This should not surprise us as one of the major achievements of humanities computing is the development of a mature discourse around the relationship between (logical) form and content for electronic texts that evolved out of discussions around markup and the Text Encoding Initiative. Recent work by McGann and Buzzetti problematizes the distinction between markup (structure) and text (content) re-proposing that markup is diacritical and that "Books are simulation machines as well, of course." (Buzzetti and McGann) In both the digital arts and humanities we are moving to working with concepts like interface, interactivity, networks, simulation, and design, that have both form and content.

3. Interactive Matters. We are, if you will, returning to an older idea of the humanities as an alternative to scholasticism and disciplinary specialization. The Italian humanists of the 15th and 16th centuries like Bruni, Alberti, Valla, and Ficino resurrected a Ciceronian ideal of interdisciplinary exchange and textual criticism that crossed then formal scholarly boundaries. They imagined a new culture of knowledge around the symposium or _convito_ that would bring together people interested in new types of questions. Likewise the emergence of interactive multimedia, from the hypertext, computer game, to interactive art, poses a challenge to disciplinary distinctions. Who should study the computer game? What sort of training does one need to create an interactive installation? What matters about interactivity?

This presentation will conclude by presenting the results of an interdisciplinary consultation (funded by SSHRC) that developed a concept paper around Interactive Matter. The iMatter project consulted widely to see if the case could be made for a strategic research creation cluster across media arts, digital humanities and game studies. One conclusion of the consultation was that our generation of "professors" are "digital immigrants" who were trained traditionally and only later crossed to digital research, while youth today are "digital natives" comfortable with digital practices as part of study across the arts and humanities. The differences encoded in disciplinary distinctions that are obvious to us are not to a generation interested in interactive media as a site for creative and playful expression. One of the things that matters, therefore, in the study of interactive works is our encounter through teaching and research with a generation for whom digital practices are not the site of interdisciplinary boundary-crossing and conflict. What can we learn about what we could be from those for whom interactivity matters?

References

Buzzetti, Dino and McGann, Jerome. "*Critical Editing in a Digital Horizon*" in _Electronic Textual Editing_ will be published by the MLA in 2006. I have quoted the preview version from http://www.tei-c.org/Activities/ETE/Preview/mcgann.xml .

McCullough, Malcolm. _Abstracting Craft; The Practiced Digital Hand_, MIT Press, 2003.

Videogames and Critical Practice: case studies and a potential future for digital humanities

Rafael FAJARDO
eMAD, SA&AH, University of Denver

I will present a case study of interactions and interventions with the Center for Interamerican Studies at the University of Texas at El Paso. These interventions have challenged rigid definitions of professional academic identities, and have been challenged by regimes that reinforce said identities. I will include successes as well as failures. These interventions occured across disciplines and across media and across bureaucratic entities. I will further introduce collaborations with collaborations with the Computer Sciences at the University of Denver where the creation of digital videogames is the integrative activity that affords and asserts a critical and humane presence in the sciences, and proposes a role for critical humanism in both game industries and marketplaces.

The interventions to be shared in this presentation were created in a situation that can best be described as a cyberpunk present which presages future conditions for digital humanities based in arts practices. This dark vision of digital practices - bereft of funding, but full of innovation - will be explored, with linkages not only to the arts and humanities, but also to contemporary economic models of the city and policy planning.

CREATING CTS COLLECTIONS

Dorothy PORTER
William DU CASSE
Jerzy W. JAROMCZYK
Neal MOORE
Ross SCAIFE
University of Kentucky

Jack MITCHELL
Stanford University

Session Contact: Dorothy Porter, Collaboratory for Research in Computing for Humanities, University of Kentucky

The Classical Text Services protocol [CTS] provides the means for coordinating and integrating XML-encoded documents on a single subject. Collaborative efforts among the Stoa Consortium and the Collaboratory for Research in Computing for Humanities at the University of Kentucky and Harvard's Center for Hellenic Studies have led to the development of several projects that seek to take advantage of CTS to provide comprehensive access to classical literary collections. Among these projects, some focus only on the textual aspects of the source materials, and the project content consists only of XML files (encoded according to the TEI Guidelines). Other projects are more elaborate and seek to link source texts to the physical artifacts which contain them, and these projects consist of TEI-XML files and digital image files.

CTS provides the means for organizing, referencing, and querying classical texts. Developed by classicists Chris Blackwell (Furman University) and Neel Smith (College of the Holy Cross), the aim of the Classical Text Services protocol is to define a network service enabling use of a distributed collection of texts according to notions that are traditional among classicists. The CTS adopts and extends the hierarchical scheme of bibliographic entities defined by the OCLC's and IFLA's Functional Requirements for Bibliographic Records, or FRBR [http://www.oclc.org/research/projects/frbr/default.htm]. FRBR describes bibliographic records in terms of a hierarchy of Works, each of which is realized through one or more Expressions, realized in turn through one or more Manifestations, realized through one or more Items. CTS implements this hierarchy using the traditional terms Work, Edition or Translation, and Exemplar, while extending the hierarchy upwards, grouping Works under a notional entity called "TextGroup'" (corresponding to authors, in the case of literary texts, or any other traditional and useful corpus, such as "Attica" for inscriptions, or "Berlin" for a published corpus of papyri). CTS also extends FRBR's hierarchy downwards, allowing identification and abstraction of citeable chunks of text (Homer, *Iliad* Book 1, Line 123), or ranges of citeable chunks (Hom.~ *Il.* 1.123-2.22). The CTS protocol allows sharing of information about texts at any level of the conceptual hierarchy, and allows retrieval of sections of an identified text at any hierarchical level supported by its scheme of citation.

In this session, we will describe the Classical Text Services protocol and explain how editors can use it to organize and query texts. We will also explain how we use CTS in one text-focused project, the Neo-Latin Colloquia project, and how we are expanding the CTS for use in the image-based Venetus A project. We end our session with a demonstration of the CTS Implementation Tool (CTS-IT), and introduce the prototype of the Network Tool for Collaborative Electronic Editing over the Internet (NeT-CEE), a tool that builds on CTS to provide support for large-scale editing projects requiring the special talents of geographically distributed individuals.

Creating a CTS Text Collection: The Neo-Latin Colloquia Project

William DU CASSE, Ross SCAIFE
University of Kentucky

Scholastic *colloquia* are didactic works from the 15th and 16th centuries designed to teach Latin to younger students through interactive listening and speaking. Accordingly, many of the *colloquia* deal with

the everyday life of schoolboys, enabling them to speak in ordinary situations using proper Latin. These dialogues, generally short and written with many idiomatic constructions, were simple enough for beginners who already possessed a basic knowledge of Latin grammar. Mastery of *sermo quotidianus* ("daily conversation") was and remains an excellent means of reaching the stage where one thinks directly in Latin rather than translating from a native language. Moreover, students who learn idiomatic Latin then read classical texts with greater facility, and their own written style improves. Writers such as Petrus Mosellanus (early sixteenth century) assert that *colloquia* offered the means for the learner to master the cultivated but familiar speech found in Cicero's letters or Terence's plays, but applied to subject matter and thoughts never treated by Cicero or Terence. Erasmus thought his contemporaries could learn the best Latin style for contemporary use by reading the best authors, but also imitating the Latin conversation of those who "spoke just as the best authors wrote."

Colloquia scholastica form a genre of largely unexplored texts that reveal much about the pedagogical practice that supported the continuing use of Latin as Europe's universal language for the educated into the Early Modern era, when it was no longer anyone's native tongue. The perpetuation of a stable language ("dead" in the argot of the linguists) based on texts only, supported by no vernacular usage, for so many centuries represents a significant and potentially illuminating linguistic phenomenon. The *colloquia* offer an untapped source from which we can learn more about the history of pedagogy during the rise of western humanism. In addition, the *colloquia* provide plentiful insights into social history. Designed to promote the use of spoken Latin for discussion of daily affairs, they reveal a great deal about the conditions and customs of life at the time when they were produced, especially about scholars, teachers, and students in and out of schools or universities, but also about conditions among the citizens, merchants, and tradesmen at large.

Over the last three years, a group at the Stoa Consortium has begun to assemble the most comprehensive collection of neo-Latin *colloquia* availably anywhere in the world, in any medium. So far we have imposed the structural markup prescribed by the Text Encoding Initiative [TEI] in its "Base Tag Set for Drama" [http://www.tei-c.org/release/doc/tei-p5-doc/html/DR.html] on over 650 *colloquia* containing over 620,000 words of text. We continue to build the collection through the addition of complex TEI markup, and we have enabled preliminary access to the collection online using the Classical Text Services protocol [CTS].

CTS allows us to access the Colloquia Collection on two levels. First, using CTS we can create a citation scheme specific to each set of *colloquia*, based on the organization of the dialogues themselves, and the internal structure of the individual dialogues. This structure is based mainly on the "Base Tag Set for Drama" as described in the TEI P5 Guidelines, so the typical citation scheme would follow the structure of a for each individual *colloquia*, containing a <sp> for each speaker in turn, containing a <p> for the spoken text. CTS thus provides us with a way to easily cite (and thus to link to) any specific point in the *colloquia* texts.

In addition, CTS also enables us to create citations for the quotations (mainly from classical authors) that appear regularly in the *colloquia*. *Colloquia* originally served as a bridge to canonical literature. Including markup to identify references, both those specifically made by the *colloquia* authors, and those identified within the text by the modern editors, will illuminate how the *colloquia* authors used the classical texts, and which classical authors and works were most influential for early modern pedagogy. Encoding these bibliographical references will also enable us in many cases to point to full texts in the Perseus Digital Library [www.perseus.tufts.edu], where users may peruse the full context of the quotations or allusions. Using the TEI markup for Bibliographic Citations and References, we identify quotations and non-quotation references that the *colloquia* make to biblical and classical texts. For example, the author Pontanus quotes and refers to classical authors (and more recent ones) after each colloquium in his *Annotationes*, a section of notes following each colloquium.

In addition to bibliographic citations, we are explicitly marking **references to specific dates, people, and place names**. These references may be to historical events or people, or to events contemporary with the writing of the *colloquia*. Using the TEI tagset for Names and Dates [http://www.tei-c.org/release/doc/tei-p5-doc/html/ND.html], we mark instances where the *colloquia* name

specific people, places, dates, and events. In the short example below, the author Vives names both a specific person and a place. TEI not only enables us to mark names as they appear in the text, but also include a regularized version to simplify both searching and reading of the text (<orig> for the original version, <reg> for the regularized version).

```
<sp>
    <speaker>Mag.</speaker>
    <p>Ubi fecisti Latinae linguae tyrocinium? Nam
    non videris mihi prave institutus.</p>
</sp>
<sp>
    <speaker>Nep.</speaker>
    <p> <placeName> <choice><orig>Brugis
    </orig><reg>Bruges</reg></choice></placeName>,
    sub <persName> <choice><orig>Joanne Theodoro
    Nervio</orig><reg>Johannes Theodorus Nervius
    </reg></choice> </persName>.</p>
</sp>
<sp>
    <speaker>Mag.</speaker>
    <p>Viro diligenti, docto, probo. <placeName>
    <choice><orig>Brugae </orig><reg>Bruges
    </reg></choice></placeName> elegantissimae:
    nisi quod pereunt in dies vitio plebis profusissimae,
    dolendum </p>
</sp>
```

(*Colloquium* 7, in Ludovicus Vives, *Linguae Latinae Exercitatio Joan. Lodov. Vivis Valentini: libellus valde doctus, et elegans, nuncque primum in lucem editus, una cum rerum, et verborum memorabilium diligentissimo indice*. Parisiis: apud Joannem Foncher et Vivantium Gaultherot, 1538, p. 298.)

In this presentation, we will discuss the reasons for building CTS support into our edition, and give examples of the encoding we use to open up the *colloquia* to querying and organization through CTS.

Using CTS for Image-Based Electronic Editions: The Venetus A Project

Jack MITCHELL
Stanford University and Dorothy Porter, University of Kentucky

The Classical Text Services protocol [CTS] enables the integration of related texts – primary text transcriptions, different editions, translations, derived works, and annotations of any of these. The MultiText H omer project (MTH) is a difficult editing project because it seeks to connect all these various types of texts relating to the *Iliad*, starting with the oldest (and, arguably, the most important) manuscript of the text, Biblioteca Nazionale Marciana, Venice, Venetus A.

In addition to the text of the *Iliad*, Venetus A includes numerous commentaries called *scholia*, which serve to describe particular aspects of the main text. The textual variants that are preserved in the scholarly commentary of the Venetus A allow us to recover some of the multiformity that was lost in the process of text fixation that gave final or near-final shape to these two monumental oral poems. Variants also give us valuable insights into the process of oral composition-in-performance. The Homer *scholia* help us reconstruct the broad diachronic dimensions inherent in the evolution of Homeric textual traditions. Different Homeric textual traditions may have been definitive at different historical moments, but no single Homeric textual version can be deemed definitive beyond its own historical context. The *scholia* to the Venetus A manuscript of the *Iliad* preserve a treasury of ancient variants and allow us the opportunity to consider many possible texts at many different stages of transmission. Through the *scholia*, we can recover both a more accurate and a more accessible picture of the fluidity inherent in the Homeric tradition, especially during the earliest stages of the text.

The *scholia*, though vital for building an understanding of the *Iliad*, have never been completely edited, or even transcribed. The reason for this may be obvious given a cursory glance at the manuscript: their appearance on the page is incredibly complex. See Figure 1 below for one small example that shows a representative section of Venetus A.

Figure 1: A slice from Venetus A, folio 111v, showing (from left to right): Marginal numbers in the Greek alpha-numeric system; marginal scholia (main collection); illuminated capital, showing the first letter in a new book; inter-marginal scholia; book title; main text of the Iliad, including interlinear notes.

The layers of text evident throughout Venetus A are perhaps best described by Thomas W. Allen in an article of 1898, the last paragraph of an article concerning the history of the creation of Venetus A [Allen]. We include it here in its entirety:

> To recapitulate the history of the MS. which we have now reconstructed; the sheets, numbered and ruled, were given out to be written. The scribe who received them wrote the text and the principal scholia in the places ruled to contain them; during the act of writing he made corrections from time to time both in the text and the scholia. This done, he apparently began the book again and wrote in the irregular space left between the scholia and the text, and between the lines of the text, other shorter scholia in a different type of hand. He took advantage of this opportunity to correct in an exhaustive manner the text he had written; he added and altered breathings, accents and apostrophes, added and corrected critical signs, and wrote above or in the inner margin corrections of words. The book, thus complete in substance, was given to the original scribe who had numbered the quires and ruled the lines; he compared it throughout with the archetype and noted on the edge of the page differences; sometimes he accompanied these with a mark to call attention; he added lines left out, and omitted scholia either in the ruled margin or the intermediate space. In a few places he explicitly refers to his authority to defend himself from corrections already made in the text by, as it would seem, the first hand on his second round. Lastly, a third person reviewed in detail the suggestions of the reviser; deleted a great number of them in favour of the reading in the text, and in other cases substituted a correction of his own. He added likewise omitted scholia and remarks of a general nature upon the context. This excessive carefulness in the preparation of the book is further seen in the numbering of the similes, the quantitative marks, and the supplements of the elisions.

For a successful edition, we need software and encoding support that will enable the editors to separate out and edit these multiple layers of text, and changes made to the manuscript over time, while specifying the relationships among those layers.

To ensure that the TEI-XML files are accessible to CTS querying, we need to ensure that the various manuscript texts (main text and *scholia*) are encoded clearly, while at the same time maintaining links among the texts and between the texts and the manuscript images. We are currently storing main text and *scholia* for each of the 24 books of the Iliad, plus introduction, in separate XML files – 50 files in all. The various *scholia* are encoded as separate divisions, one division each for marginal *scholia* (Am), intermarginal *scholia* (Aim), interior marginal *scholia* (Aint), interlinear *scholia* (Ail), book subscription (Asub), and book metrical introduction (Amet) – six divisions total. Within the divisions, groups of *scholia* that comment upon or provide alternate readings for a single lemma are grouped together as numbered segments alongside the lemma. The following example shows the encoding for the first set of marginal *scholia* in Book 4:

```
<div type=”marginal” n=”Am.4”>
    <p n=”1”>
        <seg type=”lemma”>Lemma for first marginal scholia, Book 4</seg>
```

<seg type="schol" n="1">First marginal scholion, Book 4</seg>

<seg type="schol" n="2">Second marginal scholion, Book 4</seg>

</p>

The "n" attributes provide the means by which CTS can access the *scholia* texts and build citations either from the lowest level – all the marginal *scholia* (Am – all marginal *scholia*)– to within the individual *scholion* (Am.4.1.0 – the lemma for the first marginal *scholia* of Book 4).

In addition to the XML files we have 641 image files, one for each side of the manuscript folios that survive and contain text. We are currently using the File Section and Structural Map functions of the Metadata Encoding and Transmission Standard [METS] to build indices that associate image files to the corresponding areas in the main text file, and that will associate areas of the image files with the corresponding TEI-encoded *scholia*. For this presentation, we will describe the complex textual and physical organization of the Venetus A manuscript including display of representative folios containing all of the six types of *scholia*. We will illustrate how CTS along with TEI and METS provides a robust scheme for organizing and accessing a complex image-based editing project.

Tools for Building CTS Projects: CTS-IT and NeT-CEE

Jerzy W. JAROMCZYK, Neal MOORE
University of Kentucky

For building both text editions and image-based editions, we have extended the CTS reference implementation [http://chs75.harvard.edu/projects/diginc/code/ctswebapp] to create the CTS Implementation Tool (CTS-IT). CTS-IT works with TEI files stored in an eXist XML database, providing a user-friendly interface through which an editor can upload TEI-encoded files, sort the files into groups and build a citation scheme to enable CTS querying of the data. As described in the CTS specification [http://chs75.harvard.edu/projects/diginc/specs/cts], information about the edition files is stored in the CTS Text Inventory file (TextInventory.xml), which provides an index of the files housed in the server, along with important metadata about those files.

The CTS-IT Upload function incorporates CTS metadata into the eXist upload capabilities, creating the relevant sections of the Text Inventory when storing documents in the database. CTS-IT uses the eXist API, the CTS library [http://chs75.harvard.edu/projects/diginc/code/ctslib], and the file upload functionality from Apache Commons. Functionality includes:

- Create a new textgroup in eXist and assign the textgroup a unique ID. Both the textgroup name and ID can be updated (changed) later.
- Add new texts to the textgroup
- Separate IDs for editions and translations
- Automatically validates the TEI files against DTDs or Schemas that are stored online, if they are declared in the file.
- Automatically names the uploaded file as an online node, named according to : textgroupID_projectName_projectID
- During upload, given input from the editor, CTS-IT assigns a citation scheme based on the structure of the individual files.
- Cross-check – make sure that there is a file for everything in the Text Inventory, and that every listing in the Inventory has a match in the database.
- Set eXist as a block inside the Cocoon publishing framework, so we can build on-the-fly pages based on the information stored in the database.

However, the CTS-IT provides only basic support for text-only editions. Though it is a good start, we have started development of a new tool, the Network Tool for Collaborative Electronic Editing over the Internet (NeT-CEE), which will vastly expand the functionality of CTS, allowing an editor to relate XML text files with digital images of the physical objects on which the texts are founded. NeT-CEE will allow scholars to build editions encompassing text, images, and annotations using the Extensible Markup Language (XML), the de facto standard for encoding electronic editions in the

humanities, complying with the Text Encoding Initiative (TEI) standards for markup. NeT-CEE will also support overlapping XML markup, which will occur when (for example) an edition includes markup to describe a word in a text, and that word on the page is broken between two lines.

NeT-CEE will be oriented towards collaborative electronic edition projects that bring together a number of scholars with diverse skills and interests. NeT-CEE will implement a distributed editing framework, with access control and version management systems which will allow several different editors to collaborate on an edition with different levels of access, and without fear that one editor might inadvertently overwrite another's work. Finally, since NeT-CEE will be accessible through a regular web browser it will encourage collaborative work among individuals who are geographically dispersed, and may encourage electronic editing by those many accomplished humanities scholars who are familiar with a browser view but who may be put off by regular XML editing software.Software that enables image-based collaborative editing will be applicable to countless manuscripts and papyri that have their own complex textual organizations. We will provide the means to cite not only the primary text of a document, but also the array of marginal notes and annotations that accompany it (as in Venetus A). Likewise, manuscripts of Euclid and Plato include not only marginal commentaries but tables and figures to which we wish to provide access. In addition, many historical scientific texts are already available though the CTS protocol and collaborative image-based editions could be compiled by importing new manuscript images into the existing CTS compliant texts.

We anticipate that NeT-CEE will foster the creation of scholarly works by forging partnerships between individuals and institutions, enabling them to share resources, both physical resources (in the form of texts and images) and intellectual (in the form of subject knowledge and editing experience). Because we will release NeT-CEE under an Open-Source license, it will especially promote cooperation among smaller institutions that might not have the resources to purchase expensive software. NeT-CEE will be a significant resource for scholars, but also for teachers and students, potentially encouraging collaborative projects between K-12 schools in different regions of the United States (or, indeed, around the World).

In this presentation, we will demonstrate the functionality of the CTS-IT and the prototype NeT-CEE.

[Allen] T. W. Allen. “On the Composition of Some Greek Manuscripts.” Journal of Philology, vol. XXVI, pp. 161-181, 1898.

[CTS] Classical Text Services Protocol. <http://chs75.harvard.edu/projects/diginc/techpub/cts>

[METS] Metadata Encoding and Transmission Standard <http://www.loc.gov/standards/mets/>

[TEI] TEI P5 Guidelines for Electronic Text Encoding and Interchange, edited by C.M. Sperberg-McQueen and Lou Burnard. Revised and re-edited by Syd Bauman and Lou Burnard. January 2005.

[TEXT, ANALYSIS, TOOLS].define()

G. ROCKWELL
Communication Studies and Multimedia, McMaster University

S. SINCLAIR
McMaster University

J. CHARTRAND
Open Sky Solutions, McMaster University

The unusual nomenclature of this paper's title is meant to draw attention to one of the conceptual features that has intrigued us the most in the development of the Text Analysis Portal for Research (TAPoR) : the simultaneous modularity and interdependence of the three substantives that describe our work of elaborating text analysis tools. To better understand what we are intuitively doing in developing the Portal (without always making our presuppositions explicit beforehand), and to imagine how the Portal can best fulfil its mandate as a workspace for scholars working with electronic texts and tools, we are motivated to examine text analysis tools both at the atomic and molecular level. Or, to return to the programming metaphor of the title, we wish to examine each object individually (Text, Analysis, Tools) and also as a composite object of objects (Text Analysis Tools).

The ambiguity of the method define from the title (whether it applies to the individual objects iteratively or to the collection of objects) is deliberate. The pseudo-code of this paper is polymorphous: the process of defining (that constitutes the content of the paper itself) operates on different levels and on several types of objects. Furthermore, it should be noted that the use of unquoted, capitalized words for the objects to define is deliberate: as per object-oriented convention is, these represent classes of things rather than particular instances. As such, we are not so much interested in, say, the use of tools to analyze a particular text, but rather, the particularities of texts in general as relevant to use of analysis tools. More pragmatically, we are interested in how the concepts that we take for granted in developing text analysis tools can in fact yield a rich array of useful design principles for the Portal, when examined more closely.

This paper is structured as a sequence of definitions of the relevant components in isolation, accompanied with - in increasing intensity - a discussion of how these component are transformed when combined with one another. In other words, text analysis tools are not merely the amalgam of its constituent parts, but some class of object that extends beyond them. We will conclude by outlining some of the practical consequences that these reflections might have on the next phases of developing the TAPoR Portal.

1. Text

At first glance the concept of text seems relatively easy to define, something like "a meaningful sequence of characters, or abstract symbols, that forms a structural unit." Debatable though this definition may be, it certainly allows us to identify an essential common characteristic that encompasses everything from Egyptian hieroglyphic tablets to the Gutenberg bible and even text messages that are exchanged through mobile phones. Just as importantly, it allows us to distinguish such objects from other human artefacts such as hammers, oral stories, and television shows, which do not use symbolic characters to transmit meaning within a defined scope.

However, several potential problems with this definition quickly become apparent. For instance, how do we define something as fluid and subjective as meaning? Similarly, how do we delineate as a text an object that may be structurally complex (cf. clauses, sentences, paragraphs, chapters, sections, books, volumes, etc. in prose). The latter assumes even greater significance since the structuralist and poststructuralist theorizing of intertextuality. As Roland Barthes reminds us, texts are themselves an interweaving of elements – from the etymological roots of the Latin textere for tissue – and those elements can include the most culturally diverse objects (not just other texts). This leads theorists such as Julia Kristeva to state that everything, including culture itself, is a text. Although this logical expansion of the term text is theoretically generative, such a move also ultimately renders the notion of text ineffectual, since it looses any specificity and ceases to be capable of aiding in distinguishing different types of cultural artefacts.

Potentially more interesting than the question of the scope of text is what happens to the notion of text in

a digital context. As digital textologists from Serge Lusignan (1985) and Richard Lanham (1993) to Espen Aarseth (1997) and Jerome McGann (2001) have observed, a fundamental epistemological shift occurs when moving from print to electronic textuality. In particular, although print text is composed of discreet symbols (characters) and is therefore, in a sense, already digital, the electronic medium is considerably better suited to infinite reorganizations and manipulations of those symbols; the computer makes such transformations trivial to accomplish. As a consequence, the electronic text is unstable: it is in a perpetual state of readiness to be reconfigured. And though a deformed electronic text may no longer be recognizable from its "original", it still retains associations with it through (undoable) algorithmic processes. Whereas print text can be thought of as a stable unit of meaningful characters, electronic text is better thought of as a dynamic process that encompasses several potential states for units of meaningful characters.

2. Analysis

Analysis is a classic 18th century practice worked out by John Locke and Etienne Condillac that has been adapted by humanities computing for a set of interpretative techniques that can be automated by the computer. Analysis stands in for various careful techniques of decomposing complex phenomena for the purposes of study. Digitization and computer-based tools provides us the ability to analyze large amounts of textual data quickly. This section will take three approaches to defining analysis in the context of humanities computing:

1. The difference between searching and analysis – We will look at how text analysis is different from everyday search features.
2. Five theses on analysis – We will present five theses on analysis in texual computing.
3. A reflection on analysis – We will walk through the process and results of a project to conduct analysis on analysis.

1. The difference between searching and analysis

One way into analysis is to look at what it is not. The tools of computer-assisted text analysis often resemble everyday tools. Word processors have searching tools that allow you to find a word or phrase. Such finding tools can be used as a simple text analysis environment. Likewise commercial search engines like Google do text analysis on a large scale over millions of web pages. Your word processor and Google are not, however, suited to searching large texts interactively, nor do they show you the results of a search in a way that can help you understand a literary text. Computer-assisted text analysis environments typically do three types of things beyond what the "Find" tool of a word processor might do:

i. Text analysis systems can search large texts quickly. They do this by preparing electronic indexes to the text so that the computer does not have to read sequentially through the entire text. When finding words can be done so quickly that it is "interactive", it changes how you can study the text - you can serendipitously explore without being frustrated by the slowness of the search process.

ii. Text analysis systems can conduct complex searches. Text analysis systems will often allow you to search for lists of words or for complex patterns of words, for example you can search for the cooccurence of two words or for the words before a pattern. Where you have structured text you can use the structure (typically TEI encoding) to ask questions about parts of the text.

iii. Text analysis systems can present the results in ways that suit the study of texts. Text analysis systems can display the results in a number of ways; for example, a Keyword In Context display shows you all the occurrences of the found word with one line of context as a concordance.

One can understand text analysis in the humanities as a convergence of traditions of interpretation in the humanities that evolved through print tools like the concordance with features of commercial text systems like rapid search and indexing. There is no simple history of text analysis. Instead there is a dialectic between the culture of computing and the culture of the humanities where both borrow ideas from the other. Visualization and text data mining are two new approaches that humanities computing is borrowing for analytical purposes. In the presentation we will briefly show some analytical tools borrowed from other traditions.

2. Five theses on analysis

Analysis is often understood as a set of techniques that involve the breaking apart of a complex into atomic parts

for individual study. Whether it is a complex concept that is broken down into simpler concepts or a text broken into words (or characters), analysis starts with an interruption of a continuum into parts that can be synthesized into new representations. We propose these five theses on analysis as a way of analyzing analysis:

i. Analysis is not just about breaking down an object of study into parts. Every interruption of a continuous phenomenon like a text is also a synthesis – a building up of another representation. We don't access the atomic parts by themselves, they are always represented back to us in a new synthesis of parts that pretends to be atomic.

ii. Digitization is analysis. Analysis is usually thought of as a set of practices for the study of digital texts, but the choices made in the digitization of a text and its preparation for study constrain how the text can be broken apart by the computer. To give a simple example, a digital image of a document will have different atomic parts (pixels) that are amenable to analysis than a character string. Analysis starts with decisions about what to digitize, how to digitize the what, and what formats to use. The computer can only work with the data that was input. Garbage input, garbage analyzed.

iii. The analysis is in the interface. One form in which a text is represented is through the interface of our tools, including the tools of editing and research. The relatively low resolution computer tool forces texts to be broken into facets that you scroll through, page through or navigate with hypertext links. The design of reading interfaces can thus involve a breaking down imposed by the software.

iv. Text analysis is not neutral. The act of analysis changes the phenomenon analyzed. There is the illusion of stability – that we have texts that can be studied safely without affecting the original. We will argue that there is no original electronic text, only conditions of representation that change.

v. Text analysis is in a tradition of interpretation. What matters is the conversation we have through asking questions of others and other texts. Text analysis is one way to ask questions, but it is in a tradition that involves practices that are not automated. It is a moment of the humanities, one that may be gone.

3. A reflection on analysis

This section of the paper will close with a reflection on text analysis using text analysis. We will walk through a study on text analysis using tools available in the TAPoR portal. The analysis will be reflective in the sense that it will use text analysis on texts about text analysis.

i. We will show how one can build a corpus of materials about a concept like text analysis using portal tools like the Googlizer. Just-in-time tools that build on large search engines like Google provide a way of doing conceptual analysis on the fly. This will be compared with a prepared corpus like the abstracts for the ACH/ALLC 2005 at <web.uvic.ca/hrd/achallc2005/text_analysis.htm> . The abstracts database is available from the University of Victoria web site in XML and plain text.

ii. We will show how standard analytical tools that provide word frequency lists, collocates, and repeating patterns can help one think through how the phrase "text analysis" is used on the web.

iii. We will show how a simple visualization based on cluster analysis of the corpora can suggest anomalies for further thought.

3. Tools

What is a tool in the context of humanities computing? What would defining "tool" achieve? Like many of the concepts of humanities computing, those close at hand, like "tool", are often overlooked theoretically. Tools, as Heidegger reminds us, are things at hand that you pick up and use. Work is done through the tools, without reflecting on the tools, but on the interpretative work. A good tool disappears before the interpretative work. In scholarly work, however, there are moments when the assumptions encoded in tools and techniques need to be recovered, if only to ensure that the results of interpretative practices are consistent. In this paper we will look first at four relevant definitions of tool, and then how

Working Definitions of "Tool"

Here are four candidates for what a text analysis tool is that can be illustrated by the TAPoR portal:

1. Tool as Process. An automated process for the transformation of text data. In the case of humanities

computing the process would typically be for the transformation of linguistic data or strings and it would be a process that can be executed on a computer, but need not be. The earliest text analysis tools – concording tools – took tested and useful human processes and automated them.

TAPoR encodes this definition by distinguishing between texts and tools. The distinction seem uncontroversial, until one asks just what a text is, and especially what an electronic text it. In this paper we will illustrate some problem cases encountered in the design of TAPoR.

2. Tool as Program. A utility program that implements a process (see definition 1) that is packaged in a form that can be used easily on a computer. By this definition the program is the tool, not the process. Generally a tool is not a full-blown interactive application like a word processor.

By this account MS Word would not be a tool as you can use it interactively and you can use it to do many different things even if you could use it like a tool. Grep (global regular expression print), on the other hand, is a tool that does one task efficiently. Further, the UNIX notion of tools that can be piped together evolved in the extrication of utility processes from larger environments. (See Hauben and Hauben, Netizens, chapter 9)

TAPoR treats very specific things as tools. While there are lists of tools on the web, TAPoR priviledges web services that can be used through the portal. This has the advantage that one can try the tool, but it also limits the tools available and it presumes a model of what a tool is. A problem example, XTeXT, where "one" tool is represented in the portal as many tools, will be demonstrated.

3. Tool as Technique. An intellectual technique that involves transformative or interpretative practices defined with sufficient rigor that some of the practices might be automated on the computer as processes. A technique encompasses both the human and automated practices. Even more generally one can talk about methods that might be made up of various techniques.

One way to think of tools that goes back to Engelbart's work on augmentation and to think of a tool as something that extends our capacity to do intellectual work. The tool doesn't replace us, it extends our ability to accomplish tasks. What is important is the intellectual task and the techniques that can be adapted to the task. Within the context of a task a tool can automate some part of the technique used to achieve the task.

One of the weaknesses of a tool driven project like TAPoR is that it focuses on the tools not the techniques. The intellectual techniques are taught, trained, or played with, but they cannot be fully programmed. The human transformation of internalizing a technique to the point where tools can be used transparently needs support at this juncture in humanities computing too. Some of the extensions to the portal to support training will be demonstrated.

4. Tool as Environment. An interactive environment or game in which one can run a set of transformations for a single purpose. There is obviously a grey area between an atomic tool that does one thing (if we can imagine the doing of "one" thing) and an environment that serves multiple purposes. At what point does a tool get so much functionality that it becomes an environment for processes that isn't really ONE tool but more a workbench of tools? The point, however, is that we will call an environment a tool if it is used in a context for one end. Thus Excel becomes a tool if I just use it to sort columns of text.

This distinction between tool and environment is central to the design of the TAPoR portal. The portal is a particular type of environment (a communal portal) where one has access to tools (and other things.) TAPoR encodes this distinction, which in some cases, is a draw back. The artificiality of any interface paradigm can be seen when it breaks down. A good example in the case of TAPoR is the repositories of indexed texts. Are these tools or collections of texts? (In the presentation we will review a number of these anomalies.)

What can we learn from defining tools?

What is interesting about these definitions of tools is the reflection that goes into and through tools when they are designed and used. In the second part of this paper we will step back and look generally at the rhetoric of tools. In particular we will look at a history of the software tool as a primitive in Engelbart and in the development of UNIX. This sense of a tool, as in "grep is a tool", doesn't really get at whether processes, techniques and practices are tools at all, it maintains an analogy between a class of software and other practices. We can define what a tool is, but we have to ask if "tool" is the right thing to define

in the first place. For many humanists, the word "tool" seems unsuited to humanities research. It smacks of the trades as if intellectual work was like joinery. If we look at Engelbart's language we see him using the woodworking tool analogy,

"A number of people, outside our research group here, maintain stoutly that a practical augmentation system should not require the human to have to do any computer programming--they feel that this is too specialized a capability to burden people with. Well, what that means in our eyes, if translated to a home workshop, would be like saying that you can't require the operating human to know how to adjust his tools, or set up jigs, or change drill sizes, and the like." (Engelbart, "Augmenting Human Intellect," section III.B.6)

The problem is the lack of alternatives to the tool analogy that can convey to humanists what utility programs can do. That said, we can imagine and will present an alternative analogy based on direct manipulation that would not represent text analysis as texts and tools, but as toys for manipulating texts in a game. This is the paradigm a study environment like the Ivanhoe game draws on. (See <www.speculativecomputing.org/ivanhoe/>)

An associated problem is the presumption that a tool is utilitarian - that is something used not for play, but for achieving a well defined goal, that it is a means not an end. Obviously for the tool designer the tool can be an end, but is it for user too? Users reflect on tools when they are learning them and when they break down. The experience of a new tool is not that of a known tool like a hammer, which one can pick up an use, unreflectively. A tool is not a tool at that moment of first encounter. It becomes a tool with repeated use or distraction. Humanities computing has a particular relationship with computing tools that can be seen by looking at a different discipline.

"Language is the principal - or perhaps the only - tool of the philosopher. For Wittgenstein, and for analytic philosophy in general, philosophy consists in clarifying how language can be used. The hope is that when language is used clearly, philosophical problems are found to dissolve." (Wikipedia, "Analytic philosophy" <en.wikipedia.org/wiki/Analytic_philosophy>)

Humanities computing also has a practice of clarification through tool use. Just as philosophy tries (and seems to repeatedly fail) to dissolve problems through careful language about language, humanities computing tries to engage problems through the development of computing tools, whether those tools are electronic editions, hypertexts, or text analysis programs. Encoding, in the sense of instantiating something in code, is itself a tool or practice that attempts to clarify the something sought. The problems we engage never dissolve; no tool answers our questions, that was a Wittgenstinian dream of a ladder that could be discarded. Rather, questions and problems tire and recede before new questions, like the philosophical question, what is a tool?

4. Text Analysis Tools

To this point, we have traced some of the evolution of the words "text," "analysis" and "tools" as they have transmuted over time through successive shifts in technology and practice. In contrast, the expression "text analysis tool" is a relatively recent composite term and much more closely tied to the specific contexts in which it is used (it cannot be examined generally, as we did with the other terms, because it is always already idiosyncratic for the circumstances in which it is used).

To conclude this essay, we will outline some of the ways in which the concept of text analysis tools has informed the development of the TAPoR Portal, but also how TAPoR has caused us to reconsider what we think of as text analysis tools.

References

Aarseth, E. J. (1997). Cybertext: Perspectives on Ergodic Literature. Baltimore and London: Johns Hopkins University Press.

de Condillac, E. B. (2001), Essay on the Origin of Human Knowledge, Trans. Aarsleff, H. Cambridge University Press, Cambridge.

Engelbart, D.C. (1962). "Augmenting Human Intellect: A Conceptual Framework," Summary Report AFOSR-3223 under Contract AF 49(638)-1024, SRI Project 3578 for Air Force Office of Scientific Research, Stanford Research Institute, Menlo Park, CA. Accessed online at <www.bootstrap.org/augdocs/

friedewald030402/augmentinghumanintellect/ahi62index.html>

Hanna, R (1998). "Conceptual analysis," In E. Craig (Ed.), Routledge Encyclopedia of Philosophy. London: Routledge. See <www.rep.routledge.com/article/U033>

Hauben, M. and R. Hauben (1997). Netizens: On the History and Impact of Usenet and the Internet, Wiley. A plain-text online version is available at <www.columbia.edu/~hauben/netbook/>.

Kernighan, B.W. and P.J. Plauger (1976). *Software Tools*, Addison-Wesley, Reading, MA.

Lancashire, I., J. Bradley, W. McCarty, M. Stairs, and T. R. Wooldridge (1996).*Using TACT with Electronic Texts*, The Modern Language Association of America: New York.

Lanham, R. A. (1993). *The Electronic Word: Democracy, Technology, and the Arts*. Chicago: University of Chicago Press.

Lusignan, S. (1985). Quelques réflexions sur le statut épistemologique du texte électronique. *Computers and the Humanities* 19, 209-12.

McGann, J. (2001). *Radiant Textuality : Literature after the World Wide Web*. Palgrave: New York.

JOINING UP THE DOTS: ISSUES IN INTERCONNECTING INDEPENDENT DIGITAL SCHOLARLY PROJECTS

Paul SPENCE
CCH, King's College London

John BRADLEY
King's College London

Paul VETCH
Centre for Computing in the Humanities, KCL

A characteristic of the Centre for Computing in the Humanities (CCH), King's College London, is its significant involvement with a large number of research projects that are producing digital products. At the present there are more than 30 in which our involvement is substantial, and for many of these projects our involvement stretches over a number of years.

During this time two related challenges have emerged. First, several of the projects naturally tend to group together – a user of one is likely to be interested in another as well. We have, at present, three significant groupings of this kind – a set of potentially interrelated projects about Anglo-Saxon England, a set of projects from the Classical period, and a set of Art History projects drawn from religious materials. Although the projects are done separately by different discipline specialists, there is some interest in sorting out ways that users can usefully switch from one to the other. The second challenge relates to the mix of technologies that each project uses. Some of these projects structure their materials in ways afforded by the relational model while others are using XML (primarily, of course, TEI). Eventually, perhaps, the tools available for XML will provide facilities comparable to the database engines available for the relational model, and at that point the XML model might well replace the relational one (see Bradley 2005 for some discussion about looking at XML in a relational sense), but at present this is not the case and we are finding that both our relationally-oriented projects and

our XML-oriented ones have need of both relational and XML technology to varying degrees. There are several different ways in which the two technologies can be interrelated, and at CCH we are currently working out a set of best practices for this issue ourselves.

In this session we have three papers that touch on several of these issues:

(a) The paper presented by Paul Spence describes two pieces of technology (xMod and rdb2java) that we have developed to support the presentation of XML and relational materials on the WWW. The very different natures of XML and the relational model are reflected in the very different nature of these two technologies, and several key aspects of these differences are described here.

(b) The second paper, presented by John Bradley, describes several projects from the Anglo-Saxon group of projects, and the discussion will focus on several concrete examples of the two issues mentioned above.

(c) The third paper, presented by Paul Vetch, describes some of the ways that these independent projects can be presented in, as much as possible, a consistent and unified manner over the WWW.

References

Bradley, J. (2005). Documents and Data: *Modelling Materials for Humanities Research in XML and Relational Databases. In Deegan M. (ed) Literary and Linguistic Computing*. Oxford: Oxford University Press. Vol 20 No 1. pp 133-151.

Building Web Applications That Integrate TEI XML and Relational data: xMod and rdb2java

Paul SPENCE

Both the relational database model and the document-focused XML model are regular staples in humanities computing projects, and at the Centre for Computing in the Humanities, King's College London, the two technologies have been used extensively across a large number of the thirty-plus projects that we have been involved in over the last few years. But whether for practical reasons (limited resources) or for more 'ideological' reasons (the conviction that one or other technology provides superior modelling capabilities), it seems as if relational database and document-centric XML-based approaches are often each viewed as alternatives that almost entirely preclude the use of the other.

Since both technologies tend to involve quite distinct views of how the core data may be represented, it is perhaps not surprising that a strategic choice is usually made to use one or the other across a given project, and where they do co-exist, it is often the case that development cycles are largely autonomous, with the only true points of contact being the initial strategy stage and the final presentation of the integrated data.

This was certainly the case in a number of projects involving CCH until recently, where having decided early on that the core materials required either a fundamentally 'data-centered' or a 'document-centered' approach, and then finding that the other was also necessary to represent a subset of the data, we faced significant challenges in bridging the two.

Our extensive use of TEI XML- to mark up anything from classical inscriptions; to medieval charters; to musicological bibliographies; to born digital 'presentational' websites,- has led to the creation of 'xMod', a highly modular application which can transform a heterogeneous repository of TEI XML documents into a completely finished website. Similarly, our wealth of experience in modelling humanities data (in particular prosopographical data) using relational databases has resulted in the development of an equally modular application called 'rdb2java', which facilitates the connection of a database to the web, simplifying greatly the creation of data queries, the updating of data and its final presentation. Both applications operate on common principles of separation of concerns (separating functionality from design), assist development (via the creation of a base layer of programming logic that can be easily extended) and aim to use standards-based approaches as far as possible.

In spite of the similarity in the basic objectives and

conceptual approaches behind each application, the fact is that in the first two projects where both tools were deployed,[1] the process of integration happened very late in the day, at the presentational stage. Since we have found time and time again that it is inappropriate to try to shoe-horn the entire dataset for a project into a single technology, and are facing increasing requests for features that require data to be shared between the two applications in increasingly rich and complex ways, it seemed logical to explore points of connection in a much broader sense.

One way of re-examining the parallel processes that we followed in the afore-mentioned projects was to appropriate the model proposed by Jesse James Garrett in his well-known 'Elements of user experience' diagram.[2] According to this model, the development of websites crosses five planes (strategy, scope, structure, skeleton and surface), with a dissecting line cutting down through each so that we may compare and contrast the nature of each plane according to the type of technological approach taken ('Web as software interface' or 'Web as hypertext system').

Although the model does not match our situation perfectly, we actually found the comparison extremely useful. As in the Garrett diagram, the bottom and top layers were the two parts of our development process that were integrated most closely. At the bottom ('strategy') level, project objectives were, of course, set globally, and user needs were easily assessed independently of technological approach taken. Similarly, at the top ('surface') layer, the visual design of each component had to be co-ordinated in such a way that an integrated digital publication was produced. Here there had to be some consistency between the two areas of the site, and although there were some obvious presentational differences stemming from the particular nature of each technology, it had to seem to the user as if they both formed part of a seamless whole as far as was possible.

However, in the intermediate three stages, there were some key differences. These particularly interested us, because having decided to investigate the extent to which there could be greater integration between our database and TEI XML applications, we started to ask ourselves to what degree the differing underlying models proposed by each technology would be an insurmountable obstacle in developing an integrated application that was not only seamless on the outside, but also on the inside.

Following Garrett's model, the 'web as software' path starts at the 'scope' level by defining the 'feature set' for a given set of data. The next stages are to create a structure that governs how the system responds to the user, and then to build an interface that manages these different interactions. The primary focus is on 'tasks'.

Meanwhile, the 'web as hypertext system' model begins with an analysis of how the content is to be arranged, translates this into a structural arrangement of content elements and then adds the navigational design necessary to negotiate the information structure. The primary concern here is 'information', and there is often a strong bias towards the 'document' view of data.

Leaving aside the web publication challenges, and focusing instead on the analytical paths and processes that each technology encourages, we find further differences. The process of marking up a text using TEI XML often involves adding structure to texts which already exist, thereby placing significant emphasis on the archival integrity of the source 'text', whereas in the relational database view of data, a given structure is modelled first and then populated with data. In this sense, TEI XML is often 'reporting on' or 'describing' data, whereas relational databases model, aggregate and order abstract data.

In textual markup projects, most of the decisions about technical approach and data structure are taken near the beginning of the life of the project, at the stage of document analysis. This is less true of humanities-focused database projects however- for while the process does start with some intense initial analysis, the perspective can change considerably as the project progresses since the cognitive process takes place as data is added to the project, necessitating changes to the database structure and making it more difficult to visualise the data presentation in the early stages.

Translating this to the experience of our two applications, the document-focus of xMod means that there is a much closer correlation between data representation (TEI XML markup) and its presentation, making it far easier to produce initial output using this tool than is the case with rdb2java, where more work needs to be done to display the data.

However, when it comes to searching/indexing, we find

that the hierarchical 'tree-view' model proposed by XML has a strong effect on the kinds of query that it is easy to carry out, and the relative performance of each, whereas relational data can be re-ordered and queried in more flexible ways. At one point, native XML databases seemed to provide the means for complex XML structures to be queried in an efficient manner, but in our experience, they have not so far lived up to their expectations and still fall short in terms of efficiency and scalability.

In this paper we will outline plans to build an integrated application that deals with many of these challenges, and in doing so, reconciles as far as possible the different technological perspectives that each brings to this unlikely association. At a more basic level, this includes unified design principles managed from a single point, with common CSS/XHTML components, shared libraries of styles, shared overall wireframe design and where appropriate, a single means of maintaining the central navigation system for a given project.

We will also describe some of the more complex associations possible when data is connected, shared or 'piped' between the two types of application, and explore generic ways in which we can facilitate the transmission of data to and from each in a manner that will, moreover, facilitate wider interoperability with other projects using different technological approaches.

Finally, we will discuss the ramifications of a more integrated database-TEI XML strategy for the overall project development process, with an outline of some of the technical strategies that might facilitate common development.

Footnotes

1 These were CCEDb and PASE: The Clergy of the Church of England Database, King's College London, the University of Kent at Canterbury and the University of Reading. Accessed 2005-11-11. http://www.ccedb.org.uk/. The Prosopography of Anglo-Saxon England. King's College London and Cambridge University. Accessed 2005-11-11. <http://www.pase.ac.uk/>

2 The Elements of User Experience, Jesse James Garrett website. Accessed 2005-11-11. <http://www.jjg.net/elements/pdf/elements.pdf>

References

CCH projects page. Centre for Computing in the Humanities. Accessed 2005-11-11. <http://www.cch.kcl.ac.uk/projects/>

xMod, a TEI-based publishing application. Centre for Computing in the Humanities. Accessed 2005-11-11. <http://www.cch.kcl.ac.uk/xmod/>

Garrett, Jesse James (2002). *The elements of user experience: user-centered design for the web*. New York: AIGA/New Riders.

Unity and Diversity: Finding Common Ground Among Separate Anglo-Saxon Digital Projects.

John BRADLEY

The Centre for Computing in the Humanities (King's College London) (CCH) is currently engaged with members of the Department of History at KCL, and the Department of Anglo-Saxon, Norse, and Celtic in Cambridge in the development of three digital projects related to Anglo-Saxon England. *The Prosopography of Anglo-Saxon England* (PASE) has been under development since 2000 and has recently gone online at www.pase.ac.uk. In addition, CCH has been involved more recently in two other projects with Cambridge: *AsChart* – a project investigating the creation of formal digital transcripts of Anglo-Saxon legal charters and then making them available digitally, and the *eSawyer* project, which will make available online the wealth of materials which was originally collected and published in the magisterial *Anglo-Saxon Charters: an Annotated List and Bibliography* by Professor Peter Sawyer and which is now being updated in light of more recent scholarship. Both *AsChart* and *eSawyer* are an outgrowth of the work of the British Academy and Royal Historical Society's Joint Committee on Anglo-Saxon Charters. In addition, these projects have informal connections with other projects involving CCH or Cambridge that work with materials from the Anglo-Saxon period. *LangScape* (Joy Jenkyns, Oxford) aims to make accessible information about the boundary

clauses found in legal documents that describe the borders of property, and the *Corpus of Early Medieval Coin Finds* at the Fitzwilliam Museum, Cambridge.

On one hand, these various projects should be connected to each other – they all cover materials that come from the Anglo-Saxon period in England. On the other, they involve significantly different approaches to the materials – in both a technical and a scholarly sense. *PASE* strives to record information from sources written in Anglo-Saxon times about all Anglo-Saxon individuals (great and small) and non-Anglo-Saxons important in the Anglo-Saxon world, and is strongly centred on its relational database. Aspects of its technical and conceptual framework are described, with other prosopographies in which CCH is involved, in Bradley and Short 2005. *AsChart* is in some ways a classic textual edition project, exploring the use of TEI markup for the texts of the charters that fall within its mandate. *ESawyer*, which has grown out of the both the pioneering work of Prof. Sawyer in the 1960s, and from more recent scholarship, provides bibliographic information about the charter manuscripts and editions including, for each, a standardised summary of its content, a list of manuscripts where it has been preserved, a list of printed editions and translations, and a summary of scholarly commentary that has been published. Technically, *ESawyer* data has been stored in a structured database (FileMaker), but work is now underway to use XML-based markup to represent its materials.

At first glance, the jumble of technologies and the technical and scholarly issues that the bringing together of these three projects involves might seem to present an insuperable challenge. However, the use of the XML and/or relational databases is, we believe, in the case of all three projects appropriate – *PASE* (for reasons described in Bradley and Short 2005) interprets its resources in a highly structured form, and takes advantage of the rich interconnectedness of these objects that can, at least at present, be best developed and manipulated using the relational approach. *AsChart*, as a largely classic textual edition project, needs to take advantage of the rich expressive power of TEI markup over its texts. *ESawyer* is mostly a bibliographic project and sits in-between – containing structured material for which a relational-like database approach has provided benefits, but also finding that the relational model does not entirely suit several aspects of its bibliographic materials (although neither does XML).

Scholars interested in charter texts may well wish to move between these three projects. Having found an individual of interest in one charter text, for example, a scholar might well wish to see where else, if at all, the same individual turns up in other textual sources. Alternatively, a scholar who comes to the charter materials through an individual of interest may wish to consult the *ESawyer* information on that charter to get a sense of what scholarly commentary about the text is available. Connections between the character texts are fortunately well established because the "Sawyer Number", created by Peter Sawyer in the 1960's, continues to represent a definitive identifier for the charter texts. Thus, all four of our projects here, and many others, use the Sawyer number in this way and at the document level, the connections between the different projects are well established. However, other connections are potentially more difficult. Another clear link should be between people that appear in the sources and the *PASE* prosopography. Because both the *AsChart* project and *PASE* involve the same scholars, there is general agreement about who is who in the *PASE* and *AsChart* project, and part of the task in *AsChart* markup is, as a result, to formally connect references to names in the text to appropriate individuals in PASE. However, the link to individuals in the other Anglo-Saxon projects mentioned above – in spite of the strong spirit of collaboration that has characterised our work together so far – is not so straightforward. The *Coin hoard* project grows out of a different kind of research strategy to identify individuals, and in some cases strategies to map the hoard project's prosopographical materials to *PASE*'s will present difficulties. In addition, identifiers for place and pieces of property (as possessions) figures in the *PASE* database, and ways to link between *PASE* and *LangScape* (which has place and property a central focus) still have to resolved. Furthermore, even the reading of the texts varies between *LangScape* and *AsChart* in a few significant places. Thus, there will be hard collaborative work necessary if these projects are to find ways to fully link together in ways that best benefit their respective users.

Digital collaboration between independent entities has been an important subtask in the computing world for a few years. Work on digital ontologies (which aim to develop a common vocabulary and shared significance of key ideas within a community), and web services and related technologies such as WSDL (which provide mechanisms to allow separate systems to exchange formal data) have begun to tackle some of the problems, although primarily for the business community. Our Anglo-Saxon projects

have also begun to tackle the issues of sorting out at least informally common ontologies and exploring the various formal mechanisms to query and exchange data between projects, and they have begun to experience the technical and administrative problems that similar work, elsewhere, has also experienced. In this paper we will be reporting on these issues.

References

Bradley, J. and Short S. (2005). Texts into databases: the Evolving Field of New-style Prosopography. In Deegan, M (ed) *Literary and Linguistic Computing*. Oxford: Oxford University Press. Vol 20 Suppl 1:3-24.

Corpus of Early Medieval Coin Finds. Fitzwilliam Museum, Cambridge, Available at http://www.fitzmuseum.cam.ac.uk/coins/emc/

LangScape, Language of Landscape: Reading the Anglo-Saxon Countryside. http://www.kcl.ac.uk/humanities/cch/langscape/content/index.html. The principal investigator is Joy Jenkyns (Oxford and KCL).

Prosopography of Anglo-saxon England. http://www.pase.ac.uk. The co-directors for this project are Janet Nelson FBA (King's College London) and Simon Keynes FBA (Cambridge)

Sawyer, P. (1968). *Anglo-Saxon Charters: an Annotated List and Bibliography*, London: Royal Historical Society Guides and Handbooks 8.

Erik Christensen E., Curbera F., Meredith G. and Weerawarana S. (2001). *Web Services Description Language (WSDL) 1.1*. W3C online publication at http://www.w3.org/TR/wsdl

Connecting Web Resources with Deep Hyperlinking

Paul VETCH

In the field of Humanities Computing, the idea that related resources should be interlinked not merely via the relatively superficial mechanism of the hyperlink, but more fundamentally intertwined in a more structured and meaningful way, has been espoused for many years. Indeed, one of the most rewarding aspects of work in this field is that - as projects progress - the more academics begin to understand more about humanities computing and the implications it has for their research and interests, the more imaginatively they begin to perceive connections and relationships within their source materials which were never before apparent. This process of scholarly enquiry and discovery often means that the practice of creating digital humanities resources often has to accommodate considerable change as project outcomes are remodelled and refined better to reflect the true nature of the source materials as it emerges. Of course it is not only within the context of a single project that connections arise: most projects begin with knowledge of related resources and the potential for partnerships and collaboration.

There have traditionally been two approaches to integrating separate digital resources: firstly, by supplying an overarching portal fabric to tie the resources together and present the data in a unified manner; or secondly to insist on the separation of the resources and allow superficial cross linking between them. The most successful portal environments have traditionally been found in the context of the digital library, where they are almost always focussed on information retrieval by search: OCLC's FirstSearch service, for example, provides a unified environment for querying numerous digital resources although it is not well suited to a more browsing-oriented model of usage.

Between humanities computing projects, the desire for formal relationships with related resources - expressed through data connections – certainly exists, although as yet often remains unexpressed: over and above simple hyperlinks, close integration with other projects is often simply not feasible. There are several reasons for this. Firstly the Web Services approach, as it has largely so far been seen, necessitates a rather different view of a finished project to the traditional, digital library model preferred by many funding bodies who (albeit with notable exceptions) remain attached to the idea of a 'deliverable', a unique resource with a distinct identity and presence. Secondly, even when funding does allow for more flexibility, the lingering metaphor that digital resources are somehow like giant books whose intellectual

property must be protected often precludes projects from even considering the possibility of low level connections with other resources and the considerable scope for rich and rewarding collaboration this would afford.

One possible explanation for this situation is that whilst there are established technical standards for data interchange that fit perfectly the model of distributed knowledge – viz. Web Services – there are no generic standards or agreements in place to deal with the broader implications of connecting academic resources to one another, across institutional and national boundaries. In the UK, whilst there are certainly funded initiatives looking into the establishment of such a standard, we are not close to an answer, and what work is being done is focused more on the needs of the scientific community than of the humanities. Until a standard is agreed upon, we will be left in a situation where, for the most part, projects are developed in relative isolation and we must therefore continue to try and integrate resources as best we can, after the event. Moreover the thousands of rich web resources which were developed before the advent of technologies such as WDSL cannot be ignored simply because their technical implementation is outdated.

One solution to this is to start thinking about how projects can be more seamlessly integrated at a higher level – within the presentation layer. In other words: if hyperlinking is the only viable option for linking one resource to another, how can it be improved upon as a concept? The obvious answer seems at first glance to be a simple one: using hyperlinks to target low level data, *in context*, within other sites, i.e. creating connections from points deep within one site to deep within another. Although implementation may require some sort of low level access to the target site – such as analysis of the core database or XML repository so that authority lists and links can be constructed – the cost implications would nevertheless be trivial. Whilst linking resources in this way does nothing actually to *integrate* them, each individual project remains free to maintain an independent framework of secondary and supporting material, and in fact it may be beneficial to a user to see related data in a number of different contexts - as long as the transition from one context to another is sympathetic enough to allow the data to remain intelligible in each environment.

This last clause reveals why realising a 'deep hyperlinking' solution is far from simple: allowing users to jump between resources with separate visual identities, interfaces and structures presents potentially massive problems in terms of usability and, more fundamentally, HCI. If one humanities web resource is to rely, heavily, on the content of another, yet the two must retain individual branding, how will the user perceive the relationship between them? How should this interconnection be presented? The biggest difficulty lies in the fact that, allowing for the fact that users may be sent between a number of very obviously different digital resources, at all times the visual environment must keep the user aware of where he is (not only to the extent of indicating where he is within a resource, but also which resource he is actually viewing), how he got there, and what his options are (i.e. explore the current site; return to the previous site; explore a related resource on a third site, etc). There are equally tricky subsidiary issues to consider: how should URLs be managed? Should sites share a common vocabulary, extending the scope of ontologies to include consistent labels for special pages, functions and navigational concepts across connected websites? Could digital library technologies such as OpenURL be brought to bear on the problem?

Of course, from any user's point of view, what is important is not so much *how* interconnection between two or digital resources is achieved as what the *experience* of following connections will be. Put simply, any user must to be able to follow cross references and view related resources as simply and smoothly as possible. What is needed here is some way of creating a 'seamless seam': making it obvious to the user that they have switched into another resource, whilst adhering to the HCI concepts of context and orientation.

One possible, low cost approach to this problem is to think about setting up, for each project, a special 'reception' or 'landing pad' area specifically to cater for people visiting the resource from another site, built upon middleware able to detect and reassure an itinerant visitor from another resource a) where they have come from, b) how they can get back, and c) what their options are. This is not *per se* a new concept: commercial websites have for some time used query strings in urls to post data between one another to allow a site to offer customised content specific to users coming from a certain context. The challenge from the perspective of interconnected humanities computing resources is, however, that users may well want to *return* to the

context from which they have come, and pick up exactly where they left off; equally of course they may decide that they wish to stay in the context of the 'new' resource in which they find themselves.

In this paper I shall explore these issues further, discussing how we at CCH have approached such problems across a number of related projects all oriented around the field of Anglo Saxon studies. Although we are now striving to interlink these projects to a considerable degree, this sort of functionality was never a significant part of the way they were originally conceived and so each project has been developed to have a 'standalone' existence of the type I have described here. Low-level interconnects were not possible for many of the reasons outlined above, but, because of the involvement of CCH as a key stakeholder in each project, we have been in a unique position to be able to tackle experimentally some of the difficulties with deep hyperlinks, taking users from a deep context within one site directly to a related deep context in another. I will describe the HCI strategy and technical system we are developing to allow for 'context sensitivity' in our newest projects, specifically to accommodate peripatetic users as they move between one resource and the next.

Posters

Using the OED as a Learning/research Tool in Universities

John SIMPSON

Editor of the Oxford English Dictionary

The OED is one of a range of reference resources now available to students and researchers in universities.

Since 2000, when the OED went online, the results of the dictionary's major revision programme have been published online in regular (three-monthly) instalments.

The associated search software offers extensive scope for project-based research on the English language in many disciplines: English language and literature, of course, but also much more widely.

More recently, the OED online has developed a learning resources site associated with the dictionary, which offers suggestions for different types of investigation and research. And the OED is keen to hear from researchers who are interested in contributing to this by sharing their search strategies and findings with others.

The ALLC-ACH conference provides an excellent opportunity to continue this dialogue.

What Every Digital Humanities' Scholar Should Know about Unicode: Considerations on when to Propose a Character for Unicode and When to Rely on Markup

Deborah ANDERSON

Researcher, Dept. of Linguistics, UC Berkeley

The international character encoding standard Unicode provides scholars a means to encode their texts with a widely supported standard. It is the default for the World Wide Web and plays a prominent role in the P5 version of the *TEI Guidelines* (Sperberg-McQueen and Burnard 2005). Yet even with over 97,000 characters defined, Unicode is still missing characters from various specialized fields, including characters for Byzantine Greek and Latin epigraphy, and several historic and modern minority scripts (Anderson 2003). Indeed, the *TEI Guidelines* acknowledges the challenges of being able to cover the full gamut of textual materials, by noting "there will always be a need to encode documents which use non-standard characters [i.e., which are not in Unicode] and glyphs, particularly but not exclusively in historical material" (chapter 4, Sperberg-McQueen and Burnard 2005). And in chapter 25, P5 goes on to provide guidance on how to encode letters and symbols not in Unicode with a "gaiji" module.

The trend seems to be away from proposing new characters for inclusion in the Unicode Standard, at least the Unicode Technical Committee has received relatively few requests since 2004 from digital humanities projects, with the exception of a proposal by medievalists (Everson, Haugen, et al. 2005). Since characters aren't being proposed, projects must be relying on markup with entities, employing the newly proposed "gaiji" mechanism, using a font solution (i.e., using the Private Use Area or a proprietary font with non-standard encodings), or a combination of these.

However, if texts are going to be exchanged electronically and ultimately made available to future generations of students and scholars (such as via large scale digital projects such as Open Content Alliance or as a part of online teaching materials), it might be advisable in the long run for scholars to seriously consider proposing the characters to Unicode, if they are eligible. This talk will address practical considerations digital humanists should weigh when deciding whether to pursue – or forego -- standardizing the characters in their texts.

A primary consideration is to weigh the time and effort required when formally proposing a character: The process takes two to five years, and requires an advocate to work on a proposal, be available to answer questions, and to stay involved in the process.

Other issues to consider:

* Is there broad consensus from the user community in support of the character? (Deep divisions amongst scholars will discourage the standards committees from approving a character.)
* Even if a given character is identified as needed, not all characters may be approved. Scholars may need to be flexible with the standards committees, and be open to compromise.
* Some characters are unlikely to be encoded: precomposed forms, decorations that do not appear to have semantic content, idiosyncratic letters and marks, and ligatures. Color as a feature of a character is also outside the realm of Unicode.
* Note that even after a character is approved, fonts need to be created and, in the case of complex rendering of a character (or script), users may need to wait for upgrades to rendering engines (i.e., Uniscribe) before the character can be displayed properly.

Though the onus falls on digital humanists to prove the need for the requested characters to the standards committees, the rewards of standardization can outweigh the obstacles: Because the character is part of the standard, users can expect decent rendering and display behavior from off-the-shelf software. Standardizing the characters will also make them searchable by search engines, thereby making the textual materials available to a wider audience.

The poster will include a few examples of successful character encoding proposals, which can serve as useful models for humanities text encoders, if they pursue proposing new characters.

References

Anderson, Deborah. "Unicode and Historic Scripts." *Ariadne,* Issue 37, 30 October 2003.

URL: http://www.ariadne.ac.uk/issue37/anderson/intro.html

Everson, Michael, Odd Einar Haugen, et al. "N2957: Preliminary proposal to add medievalist characters to the UCS." URL: http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2957.pdf

Sperberg-McQueen, C.M., and Lou Burnard. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Revised and re-edited. Oxford — Providence — Charlottesville — Bergen, 2005. URL: http://www.tei-c.org/release/doc/tei-p5-doc/html/

The Unicode Consortium. The Unicode Standard, Version 4.0.1, defined by: *The Unicode Standard, Version 4.0* (Reading, MA, Addison-Wesley, 2003. ISBN 0-321-18578-1), as amended by Unicode 4.0.1. URL: http://www.unicode.org/versions/Unicode4.0.1/

A Study of Buddhist Chinese- A Digital Comparative Edition of *the Bieyi za ahan jing* 別譯雜阿含經 (T.100) with English Translation

Marcus BINGENHEIMER

Chung-hwa Institute of Buddhist Studies, Taipei

The Digital Comparative Edition of the *Bieyi za ahan jing* is a project undertaken by the the Chung-hwa Institute for Buddhist Studies, Taipei (www.chibs.edu.tw) and funded by a three - year grant from the Chiang Ching-kuo Foundation of Scholarly Exchange (www.cckf.org/index-e.htm).

As Humanities Computing in Chinese is still a newcomer to the scene, this poster presentation aims to introduce a project that produces a sophisticated comparative edition of texts in ancient Asian languages using modern "western" technology (XML/TEI). We hope to discuss with other participants, learn, listen and try find ways of bridging the culture gap that exists between Humanities in the West and Asia.

The *Bieyi za ahan jing* 別譯雜阿含經 (BZA) in 16 fascicles belongs to the early Chinese Buddhist texts collectively called Ahan (Āgama) sutras 阿含經 The Ahan sutras belong to the earliest stratum of Buddhist literature. Their originals in Buddhist Sanskrit are largely lost, only a few fragments have survived. Next to the Chinese tradition only the Theravāda tradition has preserved a comprehensive set of these sutras in Pāli. For the Theravāda tradition the Nikāyas, as the Ahan sutras are called here, contain the "words of the Buddha" and therefore have been extensively studied and fully translated into English, Japanese and German. On the other hand there are extremely few translations or critical editions of the Chinese Ahan sutras.

Generally, all of the 364 short sutras contained the BZA have at least one parallel in Chinese and one Pāli parallel (with commentary). Often there are several parallels in Chinese and Pāli, sometimes even a fragment

in Buddhist Sanskrit has survived. The aim of the project is to create a digital comparative edition of the BZA, which clarifies these text-clusters. The edition will be freely available to the public. Moreover we are working on an English translation of the BZA text. Textbase for Chinese is the digital CBETA edition, for Pāli text the Vipassana Research Institute has granted us permission to use the text of the Chaṭṭha Saṅgāyana CD.

The markup of the XML files is designed according to the encoding scheme of the Text Encoding Initiative (TEI) which is transformed into HTML for the user. The parallel texts in each cluster get markup that expresses the basic dialogic structure of the content, names, differentiates between prose and verse parts, and connects them to the authoritative printed versions. The clusters are linked through a comparative catalogue. If time allows, we will add phrase-level markup for better alignment of the parallels within a text-cluster. Middleware between the source files and the user application will be eXist, an XML database.

The digital edition will help to gain a better understanding of the textual history of the BZA, especially as to the as well as of the relationship between the BZA and its Chinese, Pāli and Sanskrit parallels. This is a major contribution to the research on the formation of canonical texts in early Buddhism. By a comparison of the BZA with its Chinese parallels we will learn more about the rendition techniques of the early translators. With the help of algorithisms for authorship attribution it might even be possible to profile the ideolect of the BZA's translator(s).

The use of digital analysis for studies in Buddhist Chinese textual history is still terra nova and faces a number of problems different from European languages (incomplete character sets, no automated word-segmentation etc.). Also the delivery system based on eXist is a first for Buddhist Studies as well as Humanities Computing in Taiwan.

We hope that with the poster presentation we can start a dialogue and continue to work towards a comprehensive framework for textual studies across cultures.

Exploring Self-Advocacy Through Digital Exchange

Lorna BOSCHMAN

School of Interactive Arts and Technology, Simon Fraser University, British Columbia, Canada

lboschma@sfu.ca

The use of digital technology has become more common for those with educational or financial resources; at the same time, it has tended to marginalize communities who lack access to higher education or disposable income. This situation is especially acute for individuals with developmental disabilities*. In secondary school, typically the most advanced educational level for these learners, they are sometimes excluded from interaction with emergent information and new media technologies, and as a result, their opportunities for expression have been limited. My research addresses this profound neglect and provides pragmatic methodologies to include "self-advocates" (a designation those with developmental disabilities prefer) as authors, video directors and educators. Without the direct input of self-advocates, their embodied experiences will suffer from misguided re-interpretation by others. Computer and information technologies have allowed millions of individuals to contribute to a collective understanding of the human condition, exhibited on websites and through digital media production. My research addresses this question: how to work with marginalized individuals to achieve agency in sharing their knowledge with a wider social network. In the Prologue to his book *Collective Intelligence*, Canada Research Chair and Professor at the University of Ottawa, Pierre Lévy (1997: p.xxvii) has written: "…if we are committed to the process of collective intelligence, we will gradually create the technologies, sign systems, forms of social organization and regulation that enables us to think as a group, concentrate our intellectual and spiritual forces, and negotiate practical real-time solutions to the complex problems we must inevitably confront."

This poster presentation maps the digital exchange that

took place during the first year of This Ability Media Club, a community-based media arts program held in Burnaby, BC, east of Vancouver. Since March 2005, six to eight core participants have met weekly with two paid staff, one a documentary video-maker with experience in community arts and the other, a veteran support worker who also serves as liaison to the sponsoring agency. Both are also academic researchers associated with Canadian universities. Initially, our goals were to form a group that met regularly and to train individuals to handle the technical and conceptual challenges of media production, regardless of actual form. With the advice of stakeholders, who met regularly as an Advisory Committee, a theme was developed. The emerging filmmakers were asked to create a short video work that explored the theme: "What does community or citizenship mean to you?"

Throughout 2005, six participants developed and directed a short video project or digital story. These media works will be featured on the home page at CitizenShift, a National Film Board of Canada (NFB) website for "free range media" in April 2006 at the url <http://citizen.nfb.ca/>. This Ability Media Club is co-sponsored by the NFB, the Burnaby Association for Community Inclusion (BACI) and the United Way of the Lower Mainland, in consultation with Philia – A Dialogue on Caring Citizenship.

In the first year, the project was based on principles that Berg (2004: p.195-6) identifies with Action research: "…participation, reflection, empowerment and emancipation of people and groups interested in improving their social situation or condition." This Ability Media Club is a video variation of Photovoice. In her description of the role of Photovoice within Participatory Action Research (PAR), one of the original researchers in the field, Caroline C. Wang (1999: p.187) writes, "In line with the values that characterize PAR, photovoice integrates a citizen approach to documentary photography, the production of knowledge, and social action. As Sontag has noted, "Photographs furnish evidence.""

In the case of This Ability Media Club, my research is conducted through working with the community, engaging in an active exchange. Knowledge about digital technology is transferred, between the researchers and the group, from old members to new ones and with frequent visitors to the Club. Through their explorations in documentary video and digital storytelling, self-reflexive aspects of the lives of participants are exposed. By creating the media content, the directors negotiate another level of knowledge transfer through the CitizenShift website, one that is external to the group. Data is also being created, through transcripts of media workshops where the directors interview each other and are interviewed. These texts will be coded and individual changes in outlook and behaviour will be mapped over time. The research will be tracking words, phrases or images that could indicate a shift in the participants' relationship to media content. Do they see themselves as consumers, fans, producers or instruments of social change and does that affect content creation? The themes that emerge will be compared to the researcher's ethnographic field notes and will be viewed in the context of the media works created by the directors.

In summer 2006, the Media Club will develop a method of leading digital storytelling workshops through conducting informant design sessions. The storytelling workshops will allow self-advocate artists to engage members of the general Burnaby community in constructing short narratives that reflect their human and social relationships. The intention of this exhibition project is to reflect not the history of one community, but rather, by virtue of the social exchanges guided by the self-advocate artists, creation of a much wider narrative of the municipality's story. These additional short media works will be contributed to an exhibition (50 Years of BACI) in October 2006 at Burnaby's Shadbolt Centre for the Arts. A storytelling workshop will be designed that utilizes the leadership qualities and creative and technical skills of self-advocates in the group, based on their abilities and capacity for understanding through narrative interaction. In this way, self-advocates will not only contribute to the social capital of a local community but also to an international dialogue based on their usage of information technology.

The practical basis of my poster presentation will be informed by a more theoretical investigation into social and computational theories of collective intelligence (Lévy 1997, Rodriguez 2005), social capital (Putnam 2000), community informatics (Gurstein 2000), and the involvement of artists working within disability arts and culture (Frazee 2004). My interests are in understanding the relationship between theoretical collective intelligence or social capital, its applications within a physical community and how the process of content

creation affects participants.

Although some academics may build a social network with those most connected to them by technology and education, my work attempts to connect marginalized communities with an overall collective and social network, characterized by inclusion and engagement. Community informatics addresses this issue through promoting a "bottom-up" rather than a "top-down" approach to technological use within specific communities. This approach to technological access is based on the real needs of the community, rather than the priorities of the researcher. In the 2004 keynote address at kickstART Festival of Disability Arts and Culture, Catherine Frazee examined cultural production by artists with disabilities, highlighting their unique perspectives and gifts while carefully avoiding sentimental re-interpretations of their experiences. My poster presentation will document the shift, if any, in individual self-reflexivity in reference to the texts that the artists themselves have created. Analysis of the data will determine if they have been able to enact the term "self-advocate" through their collective actions and exhibitions.

* Developmental disabilities, in this context, are defined as an impairment of general intellectual functioning manifested before the age of 22 that may also be accompanied by physical limitations.

References

Berg, Bruce L.. (2004) *Qualitative Research Methods for the Social Sciences*. Boston: Pearson Education, Inc.

Frazee, Catherine. (2004) *Keynote address at Kickstart2 Festival of Disability Arts and Culture*. Retrieved PDF October 4, 2005 from http://www.s4dac.org/.

Gurstein, Michael [ed]. (2000). *Community informatics : enabling communities with information and communications technologies*. Hershey, PA: Idea Group Pub..

Gurstein, Michael. (2003). *Effective use: A community informatics strategy beyond the Digital Divide*. *first monday* Vol. 8, No. 12. Retrieved October 8, 2005 at http://www.firstmonday.org/issues/issue8_12/gurstein/.

Hawkins, Jeff with Blakeslee, Sandra. (2004) *On intelligence*. New York: Henry Holt and Company, LLC.

Lévy, Pierre. (1997). *Collective intelligence : mankind's emerging world in cyberspace*; translated from the French by Robert Bononno. New York : Plenum Trade.

Putnam, Robert D. Putnam. (2000) *Bowling Alone: The Collapse and Revival of American Community.* New York: Simon & Shuster.

Rodriguez, Marko A. (2005) *The Hyper-Cortex of Human Collective-Intelligence Systems*. Retrieved October 4, 2005 PDF from http://soe.ucsc.edu/~okram.

Wang, C. C. (1999) Photovoice: A Participatory Action Research Strategy Applied to Women's Health. *Journal of Women's Health* , 8, 2, 185-192.

Wang, C., & Redwood-Jones, Y. A. (2001). *Photovoice ethics. Health Education and Behavior 28(5)*, 560-572.

A Response to the B2C "Cultural Hegemony" in Humanities Computing: Pliny

John BRADLEY

King's College London

For a number of years now the WWW has acted as the dominant paradigm for the delivery of digital resources for the humanities. At King College London's CCH and indeed at other computing humanities centres with which I am familiar, a "digital project" is in almost all cases equated to a project that delivers its results to a user community over the WWW. This, of course, is natural and understandable. There is a large community of potential resource users who have access to the WWW, and are already familiar with using a browser. Because access is browser-based it is relatively easy for a humanist in his/her office to start to use them. Furthermore, the technologies that allow materials to be delivered in this way–in particular XML and XSLT – are now mature, and within the grasp of many within the humanities computing community. The downside is that the browser provides limited capabilities for interaction with the resource – the user can only view it on the screen, print it out and potentially save a copy (many browser users don't know even how to do this).

Recently these limitations have begun to hit home, and groups both within Humanities Computing (HC) and outside it have been working on strategies for tackling it. Within HC, the *TaPOR* portal is an excellent of example of the kind of thing that can be done within the WWW model. It provides an interesting working arena where users provide their own texts and select from a set of operations that can be applied against them. Improving the experience of the browser user has also been in the air recently outside of HC as well with the development of *Ajax* and other *Web 2.0* initiatives. These use sophisticated browser-based extensions to allow for a richer user experience of WWW- delivered materials.

In addition to these browser-oriented developments, there has been more radical rethinking of how interaction might be enriched from the commercial sphere. There the kind of interaction that is browser-based has been labelled "B2C" ("business to customer"). In B2C transactions the customer was seen as an individual who needs little more than choose the product, indicate how s/he wanted to pay for it, and specify where the goods where to be shipped to. When businesses wanted to sell goods to large businesses over the internet – "B2B" ("business to business") – the process became more complex: one complex computer system representing the buyer needed to interact with the computer system representing seller, and the client was no longer going to be using a browser, but instead a computing system run by the client such as the purchasing system. Out of this work arose the whole model of *web services*, and with it a host of related technologies and standards. For an overall view of developments in this area from a business perspective see the *ebXML* website which purports to enable "enterprises of any size, in any global region, to conduct business using the Internet".

It has been my contention for some time that the humanities end user needs more like B2B access to resources than B2C. After all, it has been (with people like Rosanne Potter or John B Smith) the central view of digital humanities from the very beginning that humanities computing was about allowing the researcher (the "end user") to use the strengths of the machine to engage with materials of study in new ways. Simply using the computer (via the browser) to view and print materials would surely seem to them as a betrayal of several of the most important goals of humanities computing! By applying the B2B way of thinking to humanities computing we see that we need to give end users tools that support rich access to materials made available from digital archives so that they can do interesting, independent things with them.

So, what kind of interesting thing could be done? The recent Summit on Digital Tools for the Humanities, held at the University of Virginia and sponsored in part by UVa's IATH, discussed some of these issues, and it emerged that one of areas for new tools was in the area of "scholarly interpretation". The tool envisioned at this meeting supported established scholarly practice, which was understood to be based on the act of developing a scholarly interpretation out of acts of reading and note-taking, the organising of these notes as a means to support the development of an interpretation, and then writing about the structure that emerged. During

the breakout group that discussed this issue in some detail, it was agreed that there were no existing computing tools that could support the interpretive act in this way. Interestingly, having seen the idea described during the breakout session, about half the people in the room (there were 18 in all) wanted to build something to support it, and all wanted to use it! It was also interesting to note that the approach that emerged was rather conservative in that unlike much HC thinking it focuses on supported established non-computing scholarly practice rather than opening up new approaches. This was clearly surprising to some present at the summit.

Pliny is a prototype of such a tool. It reflects some recent writing (see Bradley 2003, 2004 and Bradley and Vetch 2005) I have done on the subject of scholarly interpretation and annotation. It is also influenced by our work at CCH on the *Online Chopin Variorum Edition* (OCVE) which included some facilities for user annotation and note management. It takes up the model of a computing application that, although it interacts with resources available over the Internet, provides its user with the ability to enrich materials found there with notes that are similar to annotations that they might make in a printed resource. Furthermore, since, these notes are digital, they are potentially available for further manipulation, and *Pliny* demonstrates some of things users might want to do with their note collection. The focus is on using the machine to assist the researcher to develop his/her interpretation in a largely traditional way.

Technically, *Pliny* is built on top of *Eclipse*. This is not a coincidence. *Eclipse*, an open source software product described (at least until recently) on its website as "a kind of universal tool platform - an open extensible IDE for anything and nothing in particular", provides a rich "plugin" model for software development that supports independent tool development, but promotes interactivity between these tools. This approach would allow others to extend any tool kit they might build within Eclipse in such a way that sophisticated interaction with *Pliny* and other independently developed tools would be possible.

In my demonstration I will be showing a mature *Pliny* prototype, and explain how I believe that it can support the interaction with digital materials in ways that support the development of a scholarly interpretation of them, and this in ways that are compatible with existing widespread scholarly practice. I will show how *Eclipse* supports not only the development of this tool, but the development of a whole set of independently built tools that would provide a much richer level of interaction between them than is practical elsewhere. Finally, I will describe how *Pliny* currently interacts with existing resources over the Internet, but how resource delivery within the Humanities might change to provide a richer and more sophisticated experience by taking on more of the B2B model.

References

Bradley, J. (2003). Finding a Middle Ground between 'Determinism' and 'Aesthetic Indeterminacy: a Model for Test Analysis Tools. In Deegan M. (ed) *Literary and Linguistic Computing*. Oxford: Oxford University Press, Vol. 18 No 2. pp 185-207.

Bradley, J. (2004). "What you (fore)see is what you get: Thinking about usage paradigms for computer assisted text analysis" in *Text Technology* (forthcoming). Preprint available at http://www.kcl.ac.uk/humanities/cch/jdb/papers/FofT/index.html.

Bradley, J. and Vetch, P. (2005). "Supporting annotation as a scholarly tool: experiences from the Online Chopin Variorum Edition", a presentation given at the ACH/ALLC conference 2005, Victoria BC, Canada. http://www.cch.kcl.ac.uk/legacy/staff/jdb/papers/victoria/index.html

ebXML: Enabling a Global Electronic Market (2005). Website available at http://www.ebxml.org/

Garrett, J. J. (2005). *Ajax: A New Approach to Web Applications*. In *adaptive path* website. http://www.adaptivepath.com/publications/essays/archives/000385.php

Summit on Digital Tools for the Humanities (2005). University of Virginia, 28-30 September, 2005. http://www.iath.virginia.edu/dtsummit/

TAPoR: Text Analysis Portal for Research (2002-5). http://tapor.humanities.mcmaster.ca/home.html

Web 2.0 Conference (2005). http://www.web2con.com

Customized Video Playback Using a Standard Metadata Format

Michael BUSH

Associate Director Center for Language Studies Brigham Young University
Michael_Bush@byu.edu

Alan MELBY

Professor of Linguistics Department of Linguistics Brigham Young University
akmcvp@byu.edu

A key mantra in the early days of the digital cellular telephone revolution was "anything, anytime, anywhere." Nicolas Negroponte, director of MIT's Media Lab, modified the thrust of this declaration, coining a phrase that some would say should be the slogan of the Information Age, "nothing, nowhere, never unless it is timely, important, amusing, relevant, or capable of engaging my imagination." It is growing increasingly difficult to meet the challenge posed by this high principle, given the exponential explosion of digital media that the world faces today. Sheer volume is making it increasingly difficult to accurately identify in time and space where digital assets of interest are to be found. To find what we need (or want!), when we need it, techniques are needed to not only to represent the assets, but also to describe these in a way that facilitates storage, search, retrieval, and even playback.

The Text Encoding Initiative (TEI) and the associated international and interdisciplinary standard have addressed this challenge and made it possible for a wide variety of individuals and organizations to encode texts in such a way as to facilitate sharing of encoded texts and processing tools, thus enabling important improvements in research and teaching. Yet, the Information Age is such that it is no longer possible to rely primarily on corpora created from the written word alone. Because cultural and historical artifacts of today's society are often based on digital media other than text, it is necessary to devise standard solutions in order to ensure that all resources, which exist as all sorts of digital media, are accessible, retrievable, and useable in a wide variety of settings. Similar principles were important for TEI, and they are important for digital media today, such as DVDs and audiovisual files transmitted over the Internet.

With this wide array of distribution solutions, the problem of access now takes on new forms that exist on two levels: macro (global) and micro (local). At a macro or global level, it can be a challenge to find a particular type of video asset that conforms to some list of desired characteristics. Once a video "document" (that is, an asset) is located, however, then access takes on a micro- or local-level dimension as it becomes necessary to show only those segments of interest or to *avoid* showing portions that are not pertinent or appropriate, for whatever reason, to the need at hand. Such selective use can be problematic from many standpoints, some that are technical and some that are of a legal nature, depending on the circumstances.

Whether we are talking about access at the macro (global) level or access at the micro (local) level, access at both levels is dependent upon the availability of a descriptive mechanism that is sufficiently powerful to find the right portion of the right digital video asset. Such a descriptive mechanism, or Video Asset Description (VAD), should provide clear and searchable metadata that can be combined with display systems that are based on specialized DVD players driven by lists of video clips selected from a VAD. In either case, such systems provide access at both levels, to the video materials themselves and to specific portions of the video.

For such systems to work it is necessary to efficiently describe digital media, a need addressed by the Moving Picture Experts Group (MPEG) that has developed a standard for describing multimedia content data known as MPEG-7. In contrast with previous efforts of this committee that resulted in the standards for compressing and/or streaming digital video this effort produced an XML schema, somewhat comparable in purpose to TEI. This standard, formally named "Multimedia Content Description Interface," supports some degree of semantic interpretation that is accessible by compatible software and has been adopted by the International Standards Organization (ISO).

Our group at Brigham Young University participated in the latter phases of the MPEG-7 development process, working specifically with representatives of Motorola and NHK (the Japan Broadcasting Corporation). This

group developed an MPEG-7 "profile" (a subset of the MPEG-7 standard) that has been incorporated into Part 9 of the recently published MPEG-7 standard. Given the complexity of MPEG-7, it was necessary to develop these profiles as subsets of MPEG-7 "tools" (data element types) that cover certain functionalities.

Our interest in MPEG-7 grew out of our long-held desire to give teachers, researchers, and learners easy access to video clips in a wide variety of settings: homes, learning centers, offices, classrooms, and libraries. Indeed, we have based our work on desire to make available what is needed, when it is needed. It is clear that this principle is important not only for consumers or users of digital media, but for producers of the media, as well as libraries or online repositories where the digital media are stored. The end result is "customized video playback (CVP)" made possible by technologies built on descriptions created using standard metadata formats. CVP includes playback of a video asset under the control of a list of commands that define which segments of the video are played in what order and with which annotations the viewer can interact.

Our approach to Customized Video Playback helps achieve repeatable, customized viewing of a video program. There are basically two ways to achieve digital-technology-based Customized Video Playback. One approach is a file-based approach that requires clips to be digitized and encoded for playback, raising several challenges that have to be addressed. Particular drawbacks to this approach are (1) it is time-consuming, (2) it requires specialized skills, (3) it requires specialized, expensive, software and hardware tools, and (4) it can violate copyright law. A second approach, one for which we have developed important techniques and technologies, involves the creation of a content data model that very accurately describes the video asset of interest.

Voice Mining: A Promising New Application of Data Mining Techniques in the Humanities Domain

J. Stephen DOWNIE
M. Cameron JONES
Xiao HU

University of Illinois at Urbana-Champaign

1. INTRODUCTION

There is a growing interest by digital humanities researchers to examine the cultural computation utility of those data mining (DM) techniques that have been traditionally used by the scientific community. The scientific community has for a long time now used such techniques as Naïve Bayes, Support Vector Machines, Neural Networks and Decision Trees, etc. to build weather prediction systems, classification structures, risk management models and so on. Over the years, the scientific community has developed several DM experimenting environments that have reached such a level of maturity that one no longer needs to be especially an expert in the use of these systems. One such moderately easy-to-use DM toolkit is Weka (http://www.cs.waikato.ac.nz/ml/weka/) from the University of Waikato in Hamilton, NZ. Another is the Data-to-Knowledge (D2K)/Text-to-Knowledge (T2K) DM toolkit developed by the Automated Learning Group (ALG) at the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign (UIUC) (Downie et al 2005) (http://alg.ncsa.uiuc.edu/do/downloads/d2k). The relative maturity of DM environments like these has provided opportunities for humanities researchers to focus on the generation of new types of research questions that could be addressed by DM methods by freeing them from the previously arduous task of writing their own complicated DM algorithms. Witness the NORA project (http://www.nora.lis.uiuc.edu) whose project description states:

"In search-and-retrieval, we bring specific queries to

collections of text and get back (more or less useful) answers to those queries; by contrast, the goal of **data-mining (including text-mining)** is to produce new knowledge by exposing unanticipated similarities or differences, clustering or dispersal, co-occurrence and trends." [Emphasis is ours.]

Inspired by the NORA project's spirit of "…exposing unanticipated similarities or differences…and trends", this poster/demonstration is intended to illustrate some novel types of research questions that can be addressed through the intra-textual application of sophisticated DM techniques. More specifically, we are currently exploring the application of a variety of DM procedures to the text associated with utterances made by characters in plays. That is, we are deeming the utterances made by each character (i.e., the words they speak along with their frequency of use) to be the set of attributes that "define" that character.

We have coined the phrase "voice mining" as the rubric for this new line of computer-assisted textual inquiry. We chose "voice" to highlight that we are interested in the characters as "individuals" who, through the words they "speak", create their own individual identities and personas. We are thus adopting an anthropomorphic reading of the play text which constructs it as being analogous to the transcription of real human-to-human interactions. This anthropomorphic reading stance could be considered as a kind of extension to those traditional stylometric analyses wherein the collection of words (along with their relative frequencies) written by a given author makes up the set of attributes for describing that author. "Mining" was chosen to denote that DM techniques are the technological mode of exploration.

2. SAMPLE "VOICE MINING" RESEARCH QUESTIONS:

Character Identification: Is it possible to construct a DM generated model that can successfully identify the character which uttered a given line of play text? If yes, what are the characteristics that contribute to the construction of each character's unique voice? Even if not entirely successful, are there interesting patterns of confusion (i.e., are two or more "voices" consistently misattributed)? Do confusions suggest a kind of affinity or clustering among sub-groups of characters (perhaps providing otherwise non-obvious indications of such things as class or gender identity)?

Gender Identification: Is it possible to construct a DM generated model that can successfully identify the gender of the character which uttered a given line of play text? If so, what are the characteristics that make up the gender identities of the characters? If not, are there consistent and interesting patterns of confusion? Do the confusions suggest a possible "deliberate" subversion of gender roles?

Class/Status Identification: Is it possible to construct a DM generated model that can successfully identify the socio-economic class or status of the character which uttered a given line of play text? If so, what are the characteristics that make up the socio-economic status or class of the characters? If not, are there consistent and interesting patterns of confusion? Do the confusions suggest a possible "deliberate" subversion of class or status roles?

As one can see, it is possible to come up with many other similarly framed research questions under the voice mining paradigm. The repetitive nature of the sample questions above is deliberate and is designed to emphasize that our voice mining work is not being presented as technological magic box that provides the researcher with definitive and positive proofs. Rather, we want to stress that "voice mining" is an **exploratory** procedure and it is very important to use ones own intrinsic analytical abilities. As an exploratory procedure, prima facie "failure" to successfully model a given question should be seen as an invitation to explore the potential explanations for the "failure" because the "failed" model could, in fact, be based upon hitherto unnoticed groupings that indicate important stylistic or subversive intentions of the creator of the play's characters.

3. OUR POSTER/DEMONSTRATION

Because there is no a priori way to predict which particular DM techniques are the best at answering the kinds of questions that can be posited under the voice mining paradigm, we have limited our initial proof-of-concept work to three play texts drawn from the Project Gutenberg collection (http://www.gutenberg.org/):

Oscar Wilde: "The Importance of Being Earnest" (http://www.gutenberg.org/dirs/etext97/tiobe10.txt)

Bernard Shaw: "Pygmalion" (http://www.gutenberg.org/dirs/etext03/pygml10.txt)

Bernard Shaw: "Arms and the Man" (http://www.gutenberg.org/dirs/etext03/rmsmn10.txt)

These plays were chosen for our initial exploratory studies for they have:

a) relatively limited character sets;

b) relatively balanced representations of female and male characters; and,

c) a mix of characters by class and status.

Our poster/demonstration uses these three play texts as illustrative case studies in the application of voice mining techniques. Using examples run through both the Weka and the D2K/T2K toolkits, our poster/demonstration walks the audience step-by-step through the procedures necessary to conduct a voice mining exploration. These procedures include:

a) preprocessing of raw play text using PERL scripts to strip out extraneous text (e.g., stage directions, etc.);

b) conversion of clean text into the form needed by the DM tools (e.g., ARFF, CVS);

c) use and effects of possible attribute selection techniques;

d) selection and running of one (or more) DM algorithms; and,

interpretation of results output for signs of successful modeling and for indications of meaningful confusions. (See Figure 1 for an example output set from our "Important of Being Earnest" voice mining exploration on gender identity).

Male vs. Female		
Correctly Classified Instances	659	75.23%
Incorrectly Classified Instances	217	24.77%
Kappa statistic	0.4871	
Mean absolute error	0.2477	
Root mean squared error	0.4977	
Relative absolute error	50.49%	
Root relative squared error	100.489%	
Total Number of Instances	876	

(a) Stratified Cross-Validation Results

Class	True Positive	False Positive	Precision	Recall	F-Measure
Male	0.831	0.352	0.757	0.831	0.792
Female	0.648	0.169	0.745	0.648	0.693

(b) Detailed Accuracy By Class

T / P	Male	Female
Male	414	84
Female	133	245

(c) Classification Confusion Matrix

Figure 1. Results from a Naïve Bayes voice mining experiment constructed to explore gender identity in Wilde's "Importance of Being Earnest" using the Weka DM toolkit.

4. CONCLUSIONS AND FUTURE WORK

Our initial voice mining experiments show great promise as Figure 1 demonstrates with its strong positive results on the gender identification task. We hope this success will prod other humanities scholars to pose new and interesting questions that voice mining could help them explore. While this poster/demonstration of voice mining techniques is intended to be an illustrative proof-of-concept, we have begun work on selecting more challenging texts from which to extract character voices. We initially chose play texts as our proof-of-concept medium because each character is clearly associated with its own utterances by the convention of script writing. Book texts, for example, pose significant pre-processing challenges in that there are fewer reliable clues that consistently provide labels for "who is saying what" upon which to build DM input.

References:

Downie, J. S., Unsworth, J., Yu, B., Tcheng, D., Rockwell, G., and Ramsay, S. J. (2005). A Revolutionary Approach to Humanities Computing?: Tools Development and the D2K Data-Mining Framework. *Proceedings of the 17th Joined International Conference of ACH/ALLC*.

Delivering Course Management Technology: an English Department Evaluates Open Source and For-Profit Course Management Systems

Amy EARHART

English Dept., Texas A&M University

To meet the challenge of supporting our faculty who are interested in utilizing web based course management technology, we have tested four Course Management Systems (CMS), two open source and two for-profit. The poster will show results from a technology use survey conducted during Summer 2005 and our assessment results of four course management systems: WebCT, Turnitin.com, Moodle, and Sakai. The poster will argue for a "stair step," open source, locally controlled approach to CMS. Further, a demonstration class housed on each CMS will be available for viewing.

Background

As we ask our faculty members to incorporate technology into their teaching, we are faced with the challenge of delivering and supporting technologies that multilevel skilled users can manipulate. Further, faculty desire a wide variety of tools housed in a CMS, some simple and some complex. Given these concerns, Texas A&M University's English Department tested four CMS:

1. WebCT (webct.com)
2. Turnitin.com (turnitin.com)
3. Moodle (moodle.org)
4. Sakai (sakaiproject.org)

Several criteria impacted our evaluation of each CMS.

We have three departmental teaching needs:

1. Traditional classroom,
2. Computer classroom, and
3. Web delivered (distance) courses.

We are working within a department and university that supports technology use. Our University's Vision 2020 plan, the strategic plan of the University over the next years, includes 10 key goals, one of which is to "Increase Access to Knowledge Resources." The charge is that "wedding of communications and computer technology will, no doubt, yield the most formidable change in academe by 2002. Texas A&M University must lead the adaptation" (5). Given this charge and our corresponding survey results that note that close to 100% of our faculty would like to either use technology in the classroom or improve their technology skills for this use.

We are working with faculty who have a variety of technology skill levels. While we know that our faculty would like to use technology, the other side of the coin is that we are working with a faculty population that is not particularly advanced in ability. In a technology survey our users rated themselves as follows: Novice (29%) Intermediate (54.8%) Advanced (16.1%). Given the diversity of skill levels, we believe that one CMS will not meet all users' needs and plan to offer several choices. Because of our faculty skills, we have adopted a "stair step" approach to technology. We encourage faculty to use technology a bit at a time and, at the same time, help them to see the rewards of using technology in the classroom. Accordingly, we hope to introduce faculty to a simple CMS and, as their skills and usage grow, move them to more advanced CMS with additional tools.

We have the staff and facilities to support CMS. Our department is fortunate to have a developed technology center and staff. We have two full time computer staff, a Coordinator of Instructional Technology faculty position (myself), and two part time student technicians. Further, we have a number of servers housed in our server room on which to put a CMS. And, most importantly, we have a group that has the technological ability to mount, run and modify most open source CMS. Further, my position, Coordinator of Instructional Technology, is charged with developing and administering training

sessions, and one of our full time computer staff is charged with one on one, as needed support of faculty using technology in teaching. Therefore, we are able to provide a broader range of CMS than departments that must rely on the University for all support.

Choosing a Course Management System

The poster will show the differences between each of the evaluated course management systems using the following criteria:

- Accessibility
- Complexity
- Control (local v. central)
- Cost
- Customizable
- Design
- Tools

Findings

Our evaluations reveal the following:

1. Turnitin.com is an excellent introductory CMS. It is for-profit, but the license fee is low. Further, it has a simple interface and a small number of helpful tools, making it an excellent CMS for novice users.
2. Moodle is an open source, robust CMS of use to intermediate to advanced users. We are able to run Moodle on our server, provide training, add tools and customize the code. Further, the Moodle community continues to develop integrated tools that are beneficial to humanities users.
3. WebCT is offered to faculty through the University, but we have discouraged its use based on our test results.
4. Sakai remains an interesting CMS, but needs a bit more development before its tools are useful to most of our faculty users. TAMU is planning to enter the Sakai partnership, and we will continue to monitor the development of Sakai.

To illustrate the differences between the CMS, I will set up a sample course in each CMS for attendees to view.

References

Jadud, M. (2004). *"Considering the Alternative."* <http://www.cs-ed.org/blogs/mjadud/archives/000556.html>

Katz, S. (2005). *"Why technology matters: the humanities in the twenty-first century." Interdisciplinary Science Reviews*. 30.2: 105-118.

Palmquist, M. (2003*). "A brief history of computer support for writing centers and writing-across-the-curriculum programs." Computers and Composition*. 20.4 395-413.

Texas A&M University. (2000). *"Vision 2020: Creating a Culture of Excellence."* College Station: Texas A&M University.

Unsworth, J. (2004). "The Next Wave: Liberation Technology," *The Chronicle of Higher Education's Chronicle Review*.

Corpus Linguistic Techniques to Reveal Cypriot Dialect Information

Katerina T. FRANTZI
frantzi@rhodes.aegean.gr

Christiana LOUKAIDOU
christiana.loukaidou@gmail.com

Dept. of Mediterranean Studies
University of the Aegean
Dimokratias 1, 85100
Rhodes, Greece

The objective of this work is the creation and exploitation of a corpus consisting of traditional poems for the extraction of dialect information. We focus on the island of Cyprus. Though there have been various collections of traditional Cypriot songs and poems, they all are in hardcopy forms and as such they have been only processed manually. Language resources on minority languages are still in their infancy (McEnery et al., 2000). A corpus in electronic form however, can be processed using corpus linguistic techniques in a complete, accurate and quick way that could not be achieved manually (Biber, 1998; Ooi, 1998).

The Cypriot dialect belongs to the Eastern-Greek dialects with 18 local variations, while that and the Tsakonian dialect are the oldest Greek dialects still alive (Κοντοσόπουλος, 2001). Because of Cyprus rich history, words of various origins can be found in the Cypriot dialect: the ancient Cypriot dialect of the Achaeans, the Homer age, the Hellinistic period, the Medieval French, the Italian, the Catalan, the Arabic, the Turkish, the English. The dialect is very rich in words and expressions with ancient roots. Depending on the area, the dialect cannot be easily (if at all) understood by non-Cypriot Greeks. The dialect is mainly used in spoken language. Otherwise, it is mainly used for writing down spoken language, e.g. poems, fairytales, theatre dialogues etc. (Καρυολαίμου, 2001). Nowadays however, due to the modern way of living, the media, tourism and education, young Cypriots tend to use the common modern Greek instead of their dialect.

The Cypriot traditional poems are not simply an expression of their creators. They express the people who share the same geographical area and the same time of the poem's creation. They express their lives, feelings, hopes. They can be categorized into various types: love, satiric, immigration, history, social, religious, celebration, national and more. The language they use is simple and direct, same as that used by common people. As such, the exploitation of the poems could reveal dialect characteristics of the real language use (Χατζηιωάννου, 1999a; Χατζηιωάννου, 1999b).

We apply corpus linguistic techniques on a corpus of Cypriot traditional songs that we created to extract dialect information. We present a sample of the linguistic information extracted for the geographical areas studied. Our corpus is currently consisted of traditional poems from two geographical areas, Nicosia and Famagusta. The size of the Nicosia corpus, in terms of number of words, is 88,637, while the size of the Famagusta corpus is 82,625 words. For the comparisons between the two areas we normalize frequencies to 88,000 words. We introduce three dialect phenomena that we currently explore.

The first important dialect characteristic of the Cypriot dialect electronically explored, is the use of the archaic verbal endings "-ασιν" and "-ουσιν" (pronounced "-asin" and "-usin") for the third person in plural, past and present tense respectively, instead of the common modern Greek ones "-αν" and "-ουν" (pronounced "-an" and "-un"). For example, in the Nicosia corpus, we find the dialect word "επήρασιν" (pronounced "epirasin") instead of the common modern Greek "επήραν" (pronounced "epiran"), meaning "they took". The archaic endings survive the after-Byzantium period and are used interchangeably to the modern Greek ones, which appear at the beginning of the Middle-Age period. It is believed that the archaic endings are extensively used in some areas.

Over the total of verbs ending in "-ασιν" and "-αν", 25.11% end in "-ασιν" in Nicosia, and 43.96 in Famagusta. In the area of Famagusta, the percentage of the use of the dialectic ending "-ασιν" is almost double to that of Nicosia. The residents of Famagusta, being in a rather rural area, use the dialect without many changes. Over the total of verbs ending in "-ουσιν" and "-ουν",

25.73% end in "-ουσιν" in Nicosia, and 32.79 in Famagusta. The use of the archaic ending "-ουσιν" is also higher in Famagusta than in Nicosia. The phenomenon is explored deeper in terms of the type of verbs that prefer the archaic endings "-ουσιν" and "-ασιν" and those that prefer the common modern Greek "-ουν" and "-αν". Another parameter for future exploitation is the context analysis for both type endings, i.e. could it be that archaic endings prefer a specific syntactic of semantic environment?

A second interesting research question is on the use of the dialect negatives "εν" (pronounced "en") and "μεν" (pronounced "men"), instead of the common modern Greek "δεν" (pronounced "then") and "μην" (pronounced "min"), all meaning "not", e.g. "εν παίρνω" (pronounced "en perno") instead of "δεν παίρνω" (pronounced "den perno"), meaning "I do not take". We compare the use between "εν" and "δεν" and the use between "μεν" and "μην". Regarding "δεν", corpus processing provides us the information that in almost all the occurrences it is pronounced as "εν" in both geographical areas ("εν" is found in 92% of the cases in the Nicosia corpus, and 98% in the Famagusta corpus). Similar case we find for "μην", where the dialectic "μεν" is found in 96% of the cases in the Nicosia corpus and in 100% of the cases in the Famagusta corpus. These findings agree (as before) to that dialect characteristics are stronger in Famagusta than in Nicosia. Again, further exploitation could be on the syntactic and semantic context that the rare common Greek negatives "δεν" and "μην" tend to prefer.

Tsitakism is another interesting dialect phenomenon to explore where the "κ" ("k") is pronounced as full "τζ" ("dj"). Let us see three words found in both geographical areas: "και" (pronounced "ke") meaning "and", "καιρός" (pronounced "keros"), meaning "time" or "weather" and "κεφαλή" (pronounced "kefali") meaning "head". From our corpus we extract the information that the dialectic "τζ" ("dj") is much more common than "κ" ("k") for both geographical areas for all the three example words. The word "καιρός" is used 23 times in its dialectic form in Nicosia and only once in its common modern Greek form, while in Famagusta it only appears in its dialectic form (10 occurrences). The word "κεφαλή" is only used in its dialectic form in both geographical areas (10 occurrences in Nicosia, 1 in Famagusta). The word "and" appears in Famagusta 17 times more often in its dialectic form (379 occurences) than in its common modern Greek form (22 occurences) and in Nicosia 7 times more, i.e. 1640 on its dialectic form and 218 times on its common modern Greek form. Again, we observe the dialect phenomenon to be stronger in Famagusta.

We gave a small sample of the dialect information that can be extracted applying corpus linguistic techniques on a traditional Cypriot poems corpus. Future work focuses on the following:

- Enhancement of the corpus with more traditional poems covering more Cypriot geographical areas.
- Continuous processing of the dynamic corpus for updating the dialectic knowledge with complete and accurate information.
- Enhancement of the corpus with other forms of dialect speech.
- Continuous processing of the dynamic corpus for the extraction of extra-linguistic cultural information.
- Comparisons of the extracted dialect and extra-linguistic information with that extracted of the Dodecanese traditional songs corpus for acquiring knowledge on similarities and differences among the dialects and language idioms (Frantzi, 2005).
- Construction of an electronic Cypriot dictionary of dialect words and collocations.

References

Biber, D., Conrad, S., Reppen, R. (1998). *Corpus Linguistics - Investigating Language Structure and Use*. Cambridge University Press, Cambridge, U.K.

Frantzi, K.T. (2005). Preserving and Exploiting the Dodecanese Traditional Songs. In *Book of Extended Abstracts of the 1st South-Eastern European Digitization Initiative (SEEDI) Conference, Digital Re_Discovery of Culture (Physicality of Soul) – Playing Digital,* Ohrid, FYROM, September, 11-14 2005, pp. 41-45.

McEnery, A.M., Baker, P., Burnard, L. (2000). Corpus Resources and Minority Language Engineering. In *M. Gavrilidou, G. Carayannis, S. Markantontou,*

S. Piperidis and G. Stainhauoer (eds) Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece, 2000, pp.801-806.

Ooi, V. (1998). *Computer Corpus Lexicography*. Edinburgh University Press, Edinburgh.

Καρυολαίμου, Μ. (2001). Η Ελληνική γλώσσα στην Κύπρο. *Στο Εγκυκλοπαιδικός Οδηγός για τη γλώσσα, Α.-Φ. Χριστίδης επιστ. Υπευθ.,* Υπουργείο Παιδείας και Θρησκευμάτων, Θεσσαλονίκη, Κέντρο Ελληνικής Γλώσσας, 180-184, (in Greek).

Κοντοσόπουλος, Γ.Ν. (2001). *Διάλεκτοι και ιδιώματα της Νέας Ελληνικής*. Εκδόσεις Γρηγόρη, 3η έκδοση, Αθήνα (in Greek).

Χατζηιωάννου Κ. 1999a. *Γραμματική της Ομιλούμενης Κυπριακής Διαλέκτου με* Ετυμολογικό προσάρτημα. Εκδόσεις Ταμασός, Λευκωσία, 1999 (in Greek).

Χατζηιωάννου Κ. 1999b. *Ετυμολογικό Λεξικό της Ομιλούμενης Κυπριακής Διαλέκτου*. Εκδόσεις Ταμασός, Λευκωσία, 1999 (in Greek).

Musicology of the Future

Lorna GIBSON

Centre for Computing in the Humanities, King's College, London.

The AHRC ICT Methods Network is a new initiative which provides a national forum for the exchange and dissemination of expertise in the use of ICT (Information Communication Technology) for Arts and Humanities research. The aim of the poster is to show the Methods Network's contribution to the promotion and development of advanced use of ICT in musicology.

"We stand at the moment of opportunity"; Nick Cook's opening statement at this year's International Society for Music Information Retrieval conference (ISMIR) highlights the sense of optimism that surrounds collaborations between musicology and computer science. If we look at some of the areas being developed within computer science research centres, it is clear that this is a key topic. To name only a few; these include the research into being undertaken into intelligent sound and computing systems at the Centre for Cognition and Computing (Goldsmith's College University of London), the music and audio technology projects at the Centre for Digital Music (Queen Mary's, University of London), and the research into audio-visual media at the Lancaster Institute for Contemporary Arts (Lancaster University). Furthermore, this area of research is by no means confined to the United Kingdom; the Integrated Media Systems Centre at the University of Southern California and the German Research Centre for Artificial Intelligence (among others) are examining the study of musical performance, whilst the Music Technology Centre at McGill University (Canada) are looking into music classification software, and the Institute of Information Science in Taiwan (China) have been concerned with retrieving audio material through audio queries, to name only a few.

However, despite the fact that the tools that are being developed by such computing specialists are innovative and in many cases, highly sophisticated, they have received little response from musicologists. That is not

to say that advancements have not been made within the musicological community, far from it. Many musicologists use such electronic resources on a frequent basis for their research, and the development of resources such as Centre for the History and Analysis of Recorded Music (a major online discographic project), the Digital Archive of Medieval Music (an online resource for the study of fragments and complete manuscripts of European Medieval Polyphonic Music), and the Online Chopin Variorum Edition (online primary resource materials of Chopin's editions) show attempts at engaging with computing as part of musicological research. How musicologists use this technology within their research has yet to be seen. In general, it appears that despite the keenness of many computer scientists to advance the topic of music information retrieval, the musicological community as a whole is unsure how to take up the technological opportunities being offered and apply them to their own research.

One aspect of the AHRC ICT Methods Network's activities is to bring together these two disciplines to discuss how computer scientists can best serve musicologists, and promote and develop the use of advanced ICT methods throughout the musicological community. Whilst it is funded by a UK funding agency to promote ICT methods in the UK, the Methods Network also has an international remit and is keen to foster international partnerships and promote advanced ICT methods that are being developed outside of the UK.

The Methods Network's first major activity concerned with music is an Expert Seminar entitled "Modern Methods for Musicology: Prospects, Proposals and Realities". The aim is to bring together leading academics in music information retrieval and musicology to assess the current engagement of musicologists with information technology, consider the implications of music-computational tools for musicologists, and see what tools information scientists might offer musicologists in the future and whether these need to be adapted for use by non-technical musicians. This will be the first step to enable the Methods Network to identify current and future needs of this community, and identify strategic issues which can be addressed in follow-up workshops or other activities (which the Methods Network will support). Future directions (and activities) which arise from this event in the coming months will be included in the poster.

In summary, this poster reveals the framework for the promotion of advanced ICT research methods in music (which is one aspect of the Methods Network's activities). In doing so, it shows the importance of the Network in facilitating discussions between disciplines and in promoting collaboration, as well as the need for activities (such as seminars and training) in order to advance the use of ICT within the arts and humanities. Cook stated; "All the potential for a major disciplinary advance in musicology is there, but we've got to put in place the conditions to make it actually happen. Otherwise we could be standing at this moment of opportunity for a long time to come." The role of the Methods Network is to "make it happen" and in doing so, lay the foundations for the future musicologists.

References

Cook, Nicholas [accessed 11 November 2005]**.** *"Towards the compleat musicologist?"* http://ismir2005.ismir.net/documents/Cook-CompleatMusicologist.pdf

Crawford, Tim. [accessed 17 March 2006]. *"Music Information Retrieval and the future of Musicology"* http://www.ocve.org.uk/redist/pdf/crawford.pdf

Craig-McFeely, Julia and Marilyn Deegan. [accessed 17 March 2006]. *"Bringing the Digital Revolution to Medieval Musicology: The Digital Image Archive of Medieval Music (DIAMM)."* http://www.rlg.org/en/page.php?Page_ID=20666#article1

Fujinaga, Ichiro [accessed 17 March 2006]. *"Application of Optical Music Recognition technologies for the development of OCVE"*. http://www.music.mcgill.ca/~ich/misc/OCVE_OMR/OCVE_OMR.html

Hewlett, Walter B. and Eleanor Selfridge-Field (1991). *"Computing in Musicology, 1966-91."* Computers and the Humanities vol. 25, pp.381-392.

Wathey, Andrew, Margaret Bent, and Julia Craig-McFeely (2001). *"The Art of Virtual Restoration: Creating the Digital Archive of Medieval Music."* in "The Virtual Score: Representation, Retrieval, Restoration." Computer in Musicology vol. 12, pp. 227-240.

Towards a Union Catalogue of XML-Encoded Manuscript Descriptions

Eric HASWELL
Humanities Computing & Media Centre, University of Victoria

Matthew J. DRISCOLL
The Arnamagnæan Institute, University of Copenhagen

Claire WARWICK
School of Library, Archive and Information Studies, University College London

The recent work of the European MASTER project (1999-2001) and the Text Encoding Initiative (TEI) (Burnard & Rahtz, 2005) in developing an XML standard for encoding manuscript descriptions has provided an opportunity to explore methods of making such information more widely available through a web-based system. While traditional web database applications built on relational systems such as MySQL are well-suited for web delivery of data-centric, tabular information, they are less useful for documents with an irregular and unpredictable structure. The emergence of native XML databases as a viable storage medium has made web delivery of irregularly-structured XML easier and more efficient.

The poster discusses a prototype system developed as part of a research project for the MA in Electronic Communication and Publishing at the School of Library, Archive and Information Studies, University College London. The project, carried out at the University of Copenhagen's Arnamagnæan Institute, created a earchable, web-based catalogue of descriptions of Scandinavian manuscripts dating from the medieval period. Using PHP and the eXist native XML database, a three-tier web database application was developed. Users of the system are provided with a facility for executing queries on the database through a web form which allows complex queries to be formulated involving many different criteria. User input from a search form submission is processed by PHP into an XQuery expression which is then passed to the database. Results are returned to the user after being processed by an XSLT engine on the server. The web system is multi-lingual and places an emphasis on usability, standards-compliance and the use of open source software.

The realisation of this project is a significant first step towards the development of a comprehensive electronic tool for manuscript studies. In its current state, the system demonstrates research opportunities not previously available with manuscripts from the Arnamagnæan collection. The large number of possible search criteria, and the ability to combine these criteria in complex ways, allows researchers to assemble datasets which may otherwise have been difficult to gather. Given that there are clear limitations on the number of researchers able to physically view a manuscript due to constraints of time, funding and manuscript fragility, providing electronic access benefits researchers significantly. There also exists the possibility that researchers may discover useful and interesting information which they had previously not even considered (Driscoll, 2002).

The prototype system demonstrates a method by which manuscript repositories may undertake similar projects involving XML-encoded source material. In addition, it shows how a standardised approach to XML encoding can facilitate the integration of records from disparate repositories into a single resource, thereby creating a larger, more complete, and more useful catalogue. Indeed, this type of union catalogue was one of the primary goals of the MASTER project.

The goal of creating a unified catalogue of European medieval manuscripts may demand some measure of a standardised approach to encoding, as the ability to program query functionality is dependent on data that is structured in a similar manner across all documents being queried. Despite the availability in TEI P5 of a general tagset for encoding manuscript descriptions, the number of possible combinations of elements and different stylistic approaches to encoding present some obstacles to total integration. Surmounting the challenges imposed by encoding irregularities may be possible, however, through the use of query techniques which accommodate these differences. More work is needed to assess the implications of such an approach. It is also uncertain at this stage how feasible it is to expect the system to smoothly scale upwards as the number of documents increases. Initial indicators are positive, but a more rigorous case-study is

required.

The development of a union catalogue is therefore dependent on the availability of a technical infrastructure of sufficient flexibility and reliability. The work done in developing the prototype, and the positive results from it thus far, suggest that this is indeed possible and within reach. The eXist XML database is a viable option for storage and document management. XML documents need only to be uploaded to the database in their complete form to be added to the collection, greatly simplifying management and allowing for participating repositories to be widely dispersed geographically. A fully-functioning deployment would require a centralised server and some direct coordination of the system and its various collections, but these are logistical matters that can be readily addressed with the provision of adequate funding and, more importantly, the enthusiastic participation of manuscript repositories.

The poster will comprise of a discussion of the web resource, including examples of XML source material, the eXist database system, PHP code used to build the application, and the web interface. Further discussion will centre around the potential for involving other manuscript repositories and issues raised in this regard. Potential beneficiaries of this work might include those exploring methods of deploying XML-encoded material through a web interface, particularly if the material is not of the type that lends itself to incorporation into a relational database system, and those interested in the development of electronic tools for manuscript scholarship.

References

Burnard, Lou and Rahtz, Sebastian (2005). P5 Fascicule. [online]. Text Encoding Initiative. Available from: http://www.tei-c.org.uk/Activities/MS/FASC-ms.pdf [Accessed 27 March 2006].

Driscoll, Matthew (2002). "The MASTER Project: Defining Standards for Electronic Manuscript Catalogue Records". In: Fellows-Jensen, Gillian and Springborg, Peter (eds): Care and Conservation of Manuscripts: Proceedings of the sixth international seminar held at the Royal Library, Copenhagen 19th-20th October 2000. Copenhagen: Royal Library. pp. 8-17.

Personal Video Manager: A Tool for Navigating in Video Archives

Matti HOSIO
Mika RAUTIAINEN
Ilkka JUUSO
Ilkka HANSKI
Jukka KORTELAINEN
Matti VARANKA
Tapio SEPPÄNEN
Timo OJALA

University of Oulu, Finlande

The amount of digital information has been growing tremendously. Digital set-top-boxes are more and more common in every home. Until recently people have recorded their favourite television shows on video tapes and stored them in their bookshelves. Now they have the opportunity to record them in digital format to the hard disk of a digital receiver. This leads to growing repositories of digital video content with no annotation except the title of the recording. As the amount of data grows large enough it will be more and more difficult to find a particular video. This problem is not new. The problems were more or less the same before the digital era of today. Poorly labelled piles of VHS cassettes stored in bookshelves did not provide much help for people searching for their favourite episode of a television series recorded years ago.

The traditional way to search for a particular scene in a video stored in a large archive is to select a video and start searching from the beginning. Searching stops when the relevant position of the video is found or the end is reached. This type of search method may take a long time, especially if the user does not know the content of the video at hand well enough. Moreover, generic searches such as finding video clips with a certain theme are even more demanding using this methodology.

What we propose here is a new computer-aided way for searching and browsing through a video repository. The

video collections are processed and analyzed automatically by the computer that controls the data. Based on the analysis, the computer produces some additional data structures that will later support the search process targeted at this archive. Using this approach, the time consuming searching of videos to find something specific is no longer necessary. Actually, one does not even have to remember the title of the video one is looking for. The only information required for creating queries consists of keywords and mental imagery about the subject of interest. This kind of information is easier to remember than categorical information, such as titles of videos. In addition, the required search time using this type of methodology is only a small fraction of the time needed when using more conventional searching methods. The search concept described above is called content-based video retrieval.

A prototype of this search concept, Personal Video Manager, was developed at the MediaTeam Oulu research group [1] of the Department of Electrical and Information Engineering at the University of Oulu, Finland. It is a video search and browsing application and has been developed especially for digital set-top-boxes, as the application combines sophisticated video analysis and search software with an easy-to-use user interface. Analysis of the video material is fully automatic and thus no manual annotation is required. The user interface was designed to be simple and effective, so that the system would be usable without special training.

The research of automatic video content analysis and retrieval has years of tradition at the MediaTeam Oulu research group. The technologies developed include ones that measure visual similarities between video clips using novel relevance metrics [2][3], and efficient content-based indexing structures that utilise both automatically constructed speech transcripts and specific visual cues extracted from the videos [4]. The Personal Video Manager application makes use of these and other technologies developed during the past years.

The video material must be indexed before the search system can be used. Indexes are created by analysing the video content automatically. At first, the software segments long videos into video shots. Each shot represents one consistent camera run bounded by visual transitions. The shots are then analyzed by extracting numerical descriptions from the visual content and indexed for content-based retrieval by the search system. Text transcripts are pre-processed by a simple stemming and stop word removal and inverse document indexes are created with a temporal expansion. The resulting text indexes are matched with the extracted shot segments. A search collects and returns the shots matching the query definition. Therefore users can efficiently locate relevant spots within a time frame of seconds, even when the collections contain several hours of video.

Due to the demanding usability requirements of our prototype, a large proportion of the time spent for the system design was devoted to the user interface. The final version of the prototype provides users with two navigational methods complementary to each other. The principal method of access, which is also the more traditional one, is based on a simple keyword search. User defined query terms are compared to the text indexes of the whole video database and the most relevant video shots are returned. The system ranks the results by their relevance and displays them to the user as still images accompanied with a short piece of descriptive text. In this representation each image corresponds to a single shot stored in the archive. The displayed text is an extract of the text content in the shot. Based on the given information the user can quickly get an impression of the shot content.

Another navigational tool is based on visual similarities between video clips. For each result returned from the keyword based search, the system provides visually similar video shots based on computational similarity measurements. When a visually close match to the search need is found, it is possible to find several other relevant candidates from the visually similar results. This functionality provides an alternative way of finding relevant content, even when there are no textual content descriptions available.

By using the search methods presented above, relevant results can be obtained relatively quickly. The better the user can describe what he/she is looking for, the faster the search process is. When a relevant video shot is found, it can easily be played using the video player software provided. The user can also view the video in its entirety using conventional video playback options.

The prototype system was developed for the Oulu Expo exhibition [5], a showcase for the technological know-how of Oulu, where the public can try it out. The system contains video material consisting of over sixteen

hours of news material provided by the Finnish public service broadcasting company YLE [6]. Detailed user instructions are provided in order to make sure that all visitors of the exhibition will be able to use the system. It has been configured to collect information about how the users actually use the system. This will help to assess which of the offered navigational methods users prefer.

We strongly believe that the problem imposed by growing digital video archives will create the need for applications such as the Personal Video Manager. Such systems will probably be a standard part of future digital set-top-boxes when the required technological maturity is reached.

References

[1] MediaTeam Oulu research group. http://www.mediateam.oulu.fi/?lang=en

[2] **Ojala T, Rautiainen M, Matinmikko E & Aittola M.** (2001). *Semantic image retrieval with HSV correlograms*. Proc. 12th Scandinavian Conference on Image Analysis, Bergen, Norway, 621 – 627.

[3] **Rautiainen M & Doerman D.** (2002). *Temporal Color Correlograms for Video Retrieval*. Proc. 16th International Conference on Pattern Recognition, Quebec, Canada, 1:267 - 270.

[4] **Rautiainen M, Ojala T & Seppänen T.** (2004). *Analysing the performance of visual, concept and text features in content-based video retrieval*. Proc. 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York, NY, 197-205.

[5] Oulu Expo. http://www.tietomaa.fi/eng/nayttelyt/ouluexpo.html

[6] YLE. http://www.yle.fi/fbc/thisyle.shtml

Projet métis : passerelle entre design d'interface et création cinématographique dans le cadre des Jeux Olympiques Humanistes de Pékin 2008.

Rody KLEIN
Computer Sciences University of Savoie

Sanxing CAO
Computer sciences,
Communication University of China

Li XU
Economics, Zhejiang University

Jin CHEN
Management, Zhejiang University

Richard SMITH
communication, Simon Fraser University

Clint ROGERS
Instructional Psychology and Technology,
Brigham Young University

Nian-Shing CHEN
educational technology NSYSU Taiwan

Ghislaine CHABERT
Sciences Info Com, Université de Savoie

Todd LUBART
Psychology, CNRS Université Paris 5

Yannick GEFFROY

Metis Global virtual Network is allowing some scientists, artists, developers and other partners to shoot an advocacy advertainment movie toward Beijing Olympics 2008 while building the Camera. This metaphore means that while the creative project is evolving, the virtual team is also involved in customizing the

groupware tools and tracing the creative process in order to enhance the cooperative and individual expression. The groupware interface co-design is meant to help mediate the Metis approach to creativity (Klein, 2002), inviting to consider the universe as a continuous process of creation without totality (Levy, 1997), in which the constantly changing global context can usually be better perceived in connexion with other team members, and by allowing also the interface to be continuously adapted to an ever changing environment, along the lines of Empirical Modelling. Groupwares, very much alike adaptive forms of computing are likely to invite more contextual and sustainable innovations and may trigger more ethical concerns and deliberation than single-user-only softwares. Appropriate Groupware interface design might facilitate the very mediation of creators, developers, scientists and others within the complex field of Humanities Computing. Metis is bridging aesthetics and technology and invites to a rediscovery of the Greek Teknè and its cosmogony (Detienne, 1974; Debray, 2000).

Nous présentons dans cet article l'avancement de notre projet de recherche Métis pour favoriser à la fois la création collective mais aussi la collaboration active des informaticiens et des auteurs au sein d'une équipe virtuelle globale. Cette problématique nous amène au design d'interface d'un studio virtuel de cinéma, tout au long du projet de création pour favoriser l'expression au sein de l'équipe interculturelle.

Introduction

En novembre 2005, à 1000 jours des JO de Pékin 2008, le projet Mètis est un grand chantier exploratoire qui échappe souvent à la logique des laboratoires participants. Nous présenterons les développements du projet, les résultats ainsi que les écueils.

En 1999, le projet coordonné par Rody Klein débute à l'Université de Nice au Département Sciences InfoCom sous la supervision du psychanalyste Yannick Geffroy. Il s'agit d'analyser la définition scientifique de la créativité : "La créativité est la capacité de produire une chose qui est à la fois nouvelle (originale, inattendue) et adaptée (utile, adaptée aux contraintes) (Lubart, 1999)

Cette définition pragmatique nord américaine sera comparée à la conception européenne de la créativité qui est plutôt une "métaphore" dominée au contraire par la volonté d'expression "gratuite", détachée de toute utilité, quasiment coupée de la réalité (Joas 1999). Ce qui est un concept scientifique aux USA est une métaphore en Europe, et ce pour des raisons historiques et sociopolitiques complexes. L'étude des métaphores exige d'autres talents que l'analyse des concepts scientifiques. Elle présuppose du moins que l'on soit prêt à admettre que le phénomène cerné indirectement et à tâtons puisse faire l'objet d'une expérience réelle (Joas, 1999).

En 2001, ce constat nous a amené à explorer notre héritage culturel et à nous pencher sur les travaux concernant la Métis des Grecs (Détienne et Vernant, 1974). Ce travail nous a conduit à étudier l'œuvre du philosophe - sinologue François Jullien (1989) afin de porter un regard neuf sur l'innovation Occidentale, vue de Chine. Ces travaux ont donné lieu à la rédaction d'un mémoire (Klein, 2002) non publié. En résumé, pour permettre aux acteurs contemporains de retrouver cette prudence ancienne au sein des processus de création artistique et d'innovation technologique radical, il faudrait que grâce à la médiation d'une interface, ils puissent être sensibilisés continuellement à ce qui faisait la fibre des Humanités : la Mètis. Détienne et Vernant (1974) décrivent la Mètis des Grecs, comme une espèce d'habileté et de prudence avisée, fondée sur "la délibération en vue d'un bien". C'est la différence cruciale entre la démarche ancienne du "Technite" grec d'une part et celle d'autre part de notre ingénieur contemporain, "technicien" moderne qui a été généralement coupé d'une vision cosmologique du monde de par sa spécialisation extrême; c'est donc la différence entre « le Sage et l'Apprenti sorcier ». Cette année là, les professeurs Régis Debray (2000), Francois Jullien et Bruno Latour (1996) nous ont accueillis et ont contribué à nos orientations par leur conseils: 'vers la Chine'. La Chine où l'harmonie des individus dans la voie (Le Tao) découlait de la conscience d'appartenir à la nature: ce qui constituait la reconnaissance fondamentale de la continuité et de la connexion (corrélativité) entre toutes choses. L'innovation y était donc conçue différemment, il s'agissait d'un processus d'efficience (accommodation inclination) plutôt que d'efficacité, liant ainsi le manifeste et le latent (Jullien, 1989).

En 2003 suite a des recherches en collaboration avec

l'Ecole Nationale des Jeux vidéo et Média Interactifs basée au CNBDI d'Angoulême, se met en place un réseau international entre l'Europe, l'Amérique du Nord, la Chine et l'Afrique, afin de tenter de concrétiser un projet scientifique, technique et artistique qui permettrait d'inclure les aspects suivants :

1°) Une équipe virtuelle globale interculturelle de volontaires, composée de chercheurs en informatique et sciences humaines, de professionnels du cinéma et artistes, d'industriels et d'étudiants attirés par la recherche interdisciplinaire.

2°) Un projet de réalisation cinématographique de type Advertainement dans le cadre des Jeux Olympiques Humanistes de Pékin 2008.

3°) Un projet de Test/Conception d'une plateforme collaborative (CSCW/CSCL) permettant l'accompagnement et la Médiation artistique, scientifique, technologique et commerciale du projet Métis.

Nous faisons le point du Projet Mètis en novembre 2005 : quelques études préliminaires ont donné lieu à des publications dans les domaines du Management de l'innovation (Cao Sanxing, 2004) de l'ergonomie (Pang,2005), de la gestion des documents multimedia (Cao Sanxing, 2004, 2005) de l'éducation (Klein 2004a) et enfin des besoins technologiques des NGOs (Klein, 2005). Nous avons cherché à clarifier comment organiser une telle démarche innovante, a la fois sur le plan artistique et technologique, mais aussi aux niveaux commercial, formation et recherche. Nous avons aussi réalisé d'autres études qui décrivent les aspects concrets de la créativité collective en Chine et en Occident (XuLi et al, 2005). Enfin nous avons aussi exploré le thème des jeux Olympiques humanistes, spécifiques à Pékin 2008, en relation avec la thématique du commerce de l'Art entre esthétique et économie (Klein, 2004;Xu Li, 2005).

1. Equipe virtuelle globale : La mise en place d'une telle équipe virtuelle globale est difficile. Nous nous sommes en fait assurés d'intégrer toutes les difficultés propres à chacun des domaines : Sur le plan commercial, le projet innovant est très en amont et ne peut encore présenter de prototype, ce qui constitue un handicap à l'investissement des partenaires industriels potentiels. Sur le plan scientifique, la nature interdisciplinaire du projet n'attire qu'à moitié les laboratoires qui ne peuvent en réalité survivent que grâce à la publication d'articles très spécialisés. Par ailleurs, si la nature évolutive et artistique du projet a l'avantage d'être passionnante, elle permet cependant difficilement de savoir exactement ce que l'on veut faire et génère jusqu'à ce jour une certaine inertie quand il s'agit de demander des fonds de recherche conséquents. La question de la scientificité se pose d'autant plus que le projet acquiert une grande visibilité. Sous les feux de la rampe, la crainte des erreurs se fait de plus en plus évidente. Comme le dit Lévy-Leblond (1984) le monde des sciences et de la technologie évite de se présenter comme une errance, ce qui ne facilite pas le rapport avec les Humanités.

2. Projet Cinématographique Pékin 2008 : En ce qui concerne les Jeux Olympiques de Pékin, les publications communes (Klein,2004b; XuLi, 2005) et les accords passés avec le Laboratoire HOSC (Humanistic Olympic Study Center) nous encouragent à persévérer. Le projet d'advertainment mêle la réalisation publicitaire avec la dimension artistique. Elle peut s'appliquer aussi à des films de type « advocacy » visant à promouvoir une conception toujours plus humaniste des Jeux Olympiques où se conjuguent désintéressement de l'intention et passion des engagements sportifs, volontaire et artistique. Ne faut-il pas résister à la tentation de réduire les JO à une dimension uniquement commerciale et professionnelle? Enfin, le dispositif artistique du projet Mètis vise à décloisonner sciences, art, commerce et technologie afin de poser les questions plus larges et sociétales de la Globalisation nécessaire des cultures vers la paix et le respect de l'environnement, mais aussi de l'importance de sauvegarder la richesse de nos diversités individuelles, communautaires et écologiques. C'est un message qui passe très bien en Chine ainsi qu'en Occident : il y a donc là des réserves d'espoir et d'humanité qui pour être communiquées doivent être relayées par les TIC dont les principes technologiques et les ressorts idéologiques ne peuvent pas/ne doivent pas être seulement fondés sur la culture occidentale.

3. CSCW/CSCL : Sur le plan technologique, il est apparu difficile de commencer le projet sans outil ou tout au moins sans une connaissance partagée par les acteurs, des outils collaboratifs de base. Nous avons tenté d'utiliser sans succès l'ENT Cartable Electronique de l'Université de Savoie. Nous nous heurtons aux mêmes genres de problèmes avec le collecticiel E-Learning JoinNet de l'Université Sun Yat Sen de Taiwan.

Pourtant ce dernier est bilingue (anglais-chinois) et d'utilisation intuitive. Nous avons un administrateur de plateforme CSCL. Nous avons bien aimé le collecticiel 2.0 de Groove.net qui est d'utilisation très intuitive, mais nous n'avions aucun accès au développement. Mais d'autres problèmes se greffent là-dessus, tels que la disparité de l'accès à Internet entre les régions et les difficultés de faire tourner certains logiciels sur Windows NT version chinoise quand ils ont 'sans doute' été développés pour les versions occidentales. Saul Greenberg (Tse, 2004) nous avait avertis, de son côté qu'il avait en effet re-orienté ses recherches en CSCW sur le 'single display' pour contourner les problèmes posés par la collaboration à distance qui s'avèrent encore très importants même entre certaines régions du Canada.

Conclusion

Les travaux d'ingénierie logicielle de Pierre Lévy (CIweb) sur une théorie de l'Evolution culturelle grâce au NTIC ainsi que le nouveau langage IEML pour traquer cette évolution sur la toile, constituent une avancée notable dans l'esprit de nos travaux interdisciplinaires. Les recherches de Modélisation Empirique (EMweb) présentées l'année dernière par Mc Carty, Beynon et Russ (McCarty 2005) semblent laisser aussi envisager la possibilité de faire face avec l'informatique, à la complexité du Projet Mètis dont les besoins généralement imprévisibles seront en constante évolution à la frontière entre Humanités et Technologies. Merci à l'avance de vos suggestions et de votre aide (MètisWeb).

Références

Cao, S., Klein, R., Zheng, G., Geng, W. (2004). *Metis Global Virtual Network Toward Beijing Olympics 2008: Managing Uncertainty in Media Content Platforms. Total Innovation Management mediated by CSCW Design.* Proc. 4th Intl. Symp. on Management of Technologies. Zhejiang UniversityPress, Hangzhou, p.171-175.

Cao S., Klein R., Liu J. (2005). *Enhancing the usage pattern mining performance with temporal segmentation of QPop Increment in digital libraries.* J Zhejiang Univ SCI 6A(11):1290 1290-1296

Debray, R. (2000). *Introduction à la médiologie*. (puf).

Detienne, M., Vernant, J.P. (1974) *Les ruses de l'Intelligence, la Mètis des grecs.* (Ed. Champ Flammarion).

Joas, H. (1999). *La créativité de l'agir*, trad. De l'allemand par Pierre Rusch (éd. Cerf) Paris.

Jullien F. (1989). *Procès ou Création, Une Introduction à la Pensée des lettrés chinois*. (Des travaux / Seuil).

Klein R. (2002). *' La Mètis-pour-créer ? 'Vers l'Analyse Médiologique d'une Métaphore: La Créativité Selon la lecture de l'ouvrage de François JULLIEN (1989) : "Procès ou Création"*. Mémoire de DEA Sciences InfoCom. Université de Nice –Sophia-Antipolis, France . 223p.

Klein R. (2004a). *METIS: Chinese-Western Intercultural Approach to innovation and Global virtual teams CSCW for talent students and innovative individuals International Forum on Education Model and Policy Issues for Cultivating Talent Student* / P.1-4 , HangZhou PR China, 22- 23rd May

Klein R. (2004b). *How to Manage Global Virtual R&D cross cultural Teams using a groupware in the film industry.* Proceedings Beijing 2008 Olympic Games Symposium, HOSC, Renmin University of China. June 2004. at url: http://www.c2008.org/rendanews/english_te.asp?id=894 (accessed 10/11/2005)

Klein R., Letaief R., Carter S., Chabert G., Lasonen J., Lubart T. (2005). *CSCL for NGO's Cross Cultural Virtual Teams in Africa: An Ethiopian Children Advocacy Case Study against Exclusion and toward Facilitation of Expression, Innovation and Creativity.* Proc. ETCC-ICALT'05 Kaohsiung, Taiwan P.1037-1041.

Latour B. (1996). *Petite réflexion sur le culte moderne des dieux fétiches, Paris, Ed. Les empêcheurs de penser en rond.*

Lévy, P. (1997). *Cyberculture: Rapport au Conseil de l'Europe*. (Ed. Odile Jacob).

Lévy-Leblond J-M. (1984). *L'esprit de sel: Science, Culture, Politique*. (Points/ Nouvelle édition/ sciences).

Lubart, T.I. (1999). *Creativity across cultures*. *Handbook of creativity* (R.J. Sternberg ed.) Cambridge Press.

McCarty, W, Beynon, W.M. and Russ, S.B. (2005). *Human Computing: Modelling with Meaning*, in ACH+ALLC, 138-144

Pang N., Cao S., Schauder D., Klein R (2005). *A Hybrid Approach in the Evaluation of Usability for Multimedia Objects: Case Study of the Media Assets Management Platform for an Advertainment Production Project toward Beijing Olympics 2008* . Proceedings ICITA 05 Sidney.

Tse E. Histon J., Scott S., Greenberg S. (2004). *Avoiding Interference : How people use spatial separation and partitioning in SDG workspaces*. CSCW 2004, Chicago Vol. 6 Issue 3, P.252-261

Xu L.,Cao S.,KLEIN R., HU B.,CHEN J., JIN Y. (2005). *METIS Advertainment Movie for Beijing Olympics 2008: A Study of Cultural Economy and Cross-cultural Collaborative IT Innovation*. Proceedings Beijing 2008 Olympic Games Symposium, HOSC, Renmin University of China. June, PP. 54-68

EMweb at url: http://www.dcs.warwick.ac.uk/modelling/

CIweb - Collective Intelligence and IEML - at url : http://137.122.100.152/

MètisWeb at url: http://www.metis-global.net

Analyse d'un extrait du roman de Pieyre de Mandiargues, *La Motocyclette*, assistée par le logiciel de statistique lexicale TACT

Caroline LEBREC

French, University of Toronto

Dans le cadre d'un séminaire de théorie littéraire, nous avons choisi de faire une étude d'un texte, en utilisant l'approche de la stylostatistique. Notre propos a été d'observer ce qu'une étude du *vocabulaire* pouvait apporter à la compréhension d'un texte. Nous entendons le terme *vocabulaire* selon la définition de Jean-Marie Viprey *(Analyses textuelles et hypertextuelles des* Fleurs du mal) dont nous rappelons ici la terminologie :

> *Nous appellerons 'vocable' tout 'lexème' actualisé (employé) dans l''énoncé' observé et 'vocabulaire' le système où s'organisent les 'vocables' (autrement dit, le niveau lexical de l'énoncé). (65)*

Le choix de notre corpus littéraire s'est tourné vers le roman, car ce genre nous a semblé moins fréquenté par les études statistiques, aux contraires des genres de la poésie et du théâtre. Toujours selon Viprey :

> *Lire un poème particulier est – presque toujours la manière "normale" d'entrer dans un recueil (ce qui, au passage, est propre aux genres poétiques et n'est guère vrai du roman). Il est donc également assez conforme aux pratiques du texte poétique de s'emparer "comme au hasard", au cours d'un poème, d'un fil d'Ariane dont on ne sait pas immédiatement où il nous conduira, et c'est bien sûr le charme possible de cette lecture,* hypertextuelle *par excellence. Au point que c'est sans doute à la poésie, parmi les genres "traditionnels", que l'*hypertexte *est le plus directement utile. (63)*

Afin de suivre l'approche poétique de Viprey, nous choisissons de rompre la linéarité du récit, afin d'en étudier un extrait, plutôt que de l'étudier dans son entité.

Nous entendons par récit un assemblage structuré de fragments, tout comme un recueil de poésie est un ensemble structuré de poèmes. Ceci signifie également que nous employons l'aspect poétique de la prose de Mandiargues pour indiquer ce que Gabriel Saad appelle une *construction* et un *exercice d'écriture :*

> *À partir d'une couleur, le blanc, la prose de Mandiargues prolifère. Et c'est ainsi qu'elle devient poétique, car cette prolifération est non seulement une construction, comme nous l'avons montré dans notre analyse, mais aussi un exercice d'écriture qui consiste à donner à certains mots une valeur particulière, propre au texte dans lequel ils sont employés. Il ne s'agit donc nullement d'une dérive sémantique mais d'un jeu dont les règles sont aussi rigoureuses que celles de la peinture ou de la poésie.* (« La prose d'André Pieyre de Mandiargues : ut pictura poesis », p122)

Si Saad étudie la *combinatoire de forme et de sens* du style de Mandiargues dans une nouvelle, nous nous demandons si un fragment d'un récit de roman amène aux mêmes aspects stylistiques?

Nous étudions les occurrences lexicales de ce que Saad appelle le *fonctionnement singulier de l'écriture* de Mandiargues (113) : d'un côté, la pluralité des voix narratives (narrateur + personnage) et d'un autre côté, la *définition d'une 'palette'* (113) de couleurs reliant les thèmes de la mort et de l'érotisme ((*L'union étroite des deux thèmes : Eros et la mort*, 116).

Par conséquent, nous avons choisi d'indexer un extrait de *La Motocyclette* (Pieyre de Mandiargues, 1963) qui inclut en même temps l'aspect stylistique de la '*palette*' et l'aspect polyphonique des voix narratives. La scène choisie se trouve aux pages 118, 119 et 120 du roman (Gallimard, coll. L'Imaginaire, 1999). Dans cet extrait, Rebecca (le personnage principal) s'arrête dans une station essence pour remplir le réservoir de sa moto. Observant à distance les couleurs et les formes de ce qui est devant elle, le narrateur décrit son moment de panique face à la vision de cette station et de son pompiste. La scène se termine par la fuite : elle ouvre les gaz et n'a pas rempli son réservoir.

Le récit est pris en charge par un narrateur omniscient qui semble se dédoubler entre le récit de l'introspection de Rebecca et ses propres commentaires, qui sont marqués typographiquement par la présence de virgules ou de parenthèses. (Rappelons que l'étude de Saad avait déjà noté que la polyphonie narrative de Mandiargues était marquée typographiquement).

La méthodologie de Saad est d'abord statistique puis stylistique. Nous choisissons d'inverser le processus. En effet, nous voulons que ce soient les résultats de l'indexation du texte qui nous ramène à notre hypothèse de lecture textuelle : y a-t-il un changement lexicologique quand il y a un changement de voix narrative ?

Nous fournissons trois résultats de notre indexation de corpus.

1. La '*palette*' de Rebecca et son occurrence dans l'extrait.
2. En second lieu, la '*palette*' du narrateur et son occurrence dans l'extrait.
3. La lecture hypertextuelle vérifie si les changements lexicologiques suivent une règle générique dans les marques du genre féminin (substantifs, adjectifs et causes déterminantes).

Notre objectif est de voir si le changement de lexique peut prouver l'omniprésence du genre féminin dans le vocabulaire employé par le narrateur et donc si l'aspect polyphonique du récit est un artifice utilisé par Mandiargues pour donner un aspect androgyne au récit du narrateur.

Pour cette étude, nous avons utilisé le logiciel de traitement de données lexicales TACT, créé à l'université de Toronto par John Bradley dans les années 80. L'interface en français a été écrite par les professeurs Russell Wooldridge (Université de Toronto) et Émilie Devriendt (École Normale Supérieure de Paris). Elle est disponible en ligne sur le site du département français de l'Université de Toronto.

Nous avons utilisé les ressources MAKEBASE et USEBASE pour réaliser l'indexation de note extrait et TACTSTAT pour obtenir des résultats.

Le logiciel lui-même étant en anglais, nous avons dû faire quelques ajustements techniques concernant la non équivalence de la langue anglaise et française (par exemple la non prise en compte de accents qui peut changer la forme des '*vocables*' indexés et donc sa catégorie grammaticale).

Notre poster souligne le lien entre la statistique et son champ théorique : la stylistique.

Analysis of an Extract of the Novel of Pieyre de Mandiargues, *La Motocyclette,* Assisted by the Lexical Statistics Software TACT

Within the framework of a literary seminar of theory, we choose to study a text using the stylostatistical approach. Our intention is to observe what a study of *vocabulaire* could bring to the stylistical comprehension of a text. We define the term " vocabulaire" according to the definition of Jean-Marie Viprey *(Analyses textuelles et hypertextuelles des* Fleurs du mal):

> *Nous appellerons 'vocable' tout 'lexème' actualisé (employé) dans l''énoncé' observé et 'vocabulaire' le système où s'organisent les 'vocables' (autrement dit, le niveau lexical de l'énoncé).* (65)

The choice of our literary corpus is the novel as this literary genre has received less attention by statistical studies, which seem to prefer studies on poetry and theatre genres. Still, according to Viprey:

> *Lire un poème particulier est – presque toujours la manière "normale" d'entrer dans un recueil (ce qui, au passage, est propre aux genres poétiques et n'est guère vrai du roman). Il est donc également assez conforme aux pratiques du texte poétique de s'emparer "comme au hasard", au cours d'un poème, d'un fil d'Ariane dont on ne sait pas immédiatement où il nous conduira, et c'est bien sûr le charme possible de cette lecture,* hypertextuelle *par excellence. Au point que c'est sans doute à la poésie, parmi les genres "traditionnels", que l'*hypertexte *est le plus directement utile. (63)*

In order to follow Viprey's poetic approach, we choose to break the linearity of a novel's account and study an extract rather than studying it in its entity. Therefore, our corpus will be a scene extracted from the linearity of the account. We understand a novel as being a structural assembly of fragments, like a collection of poetry is a structural assembly of poems. This means also that we use the poetic aspect of Mandiargues' prose to reveal what Gabriel Saad calls *une construction* and *un exercice d'écriture* :

> *À partir d'une couleur, le blanc, la prose de Mandiargues prolifère. Et c'est ainsi qu'elle devient poétique, car cette prolifération est non seulement une construction, comme nous l'avons montré dans notre analyse, mais aussi un exercice d'écriture qui consiste à donner à certains mots une valeur particulière, propre au texte dans lequel ils sont employés. Il ne s'agit donc nullement d'une dérive sémantique mais d'un jeu dont les règles sont aussi rigoureuses que celles de la peinture ou de la poésie.* ((« La prose d'André Pieyre de Mandiargues : ut pictura poesis », p122)

If Saad is studying the *combinatoire de forme et de sens* of Mandiargues' style in a short story, we wonder if a fragment of a novel's account can also lead us to the same stylistical observations?

We study the lexical occurrences serving what Saad calls *le fonctionnement singulier de l'écriture de Mandiargues* (113): in one hand, the plurality of the narrative voices (narrator + character), and on the other hand, *la définition d'une 'palette'* (113) of colors connecting the topics of death and erotism (*L'union étroite des deux thèmes : Eros et la mort*, 116).

Therefore, we choose to index an extract of *La Motocyclette* (Pieyre de Mandiargues, 1963) which includes both this 'palette' aspect and the polyphonic aspect. The selected scene is on pages 118, 119 and 120 (Gallimard, coll. L'Imaginaire, 1999). In this extract, Rebecca (the main character) stops in a gasoline station to fill up her motorcycle. Observing the colours and forms of what is in front of her, the narrator describes her moment of panic regarding this place and its pump assistant. The scene ends with an escape: she starts the bike yet does not fill up the tank.

The account is recited by an omniscient narrator who seems to duplicate himself between the introspections of Rebecca and his own perceptions. His comments are indicated typographically by the presence of commas and brackets. (Let us recall that the study of Saad had already noted that the narrative polyphony in Mandiargues' style is typographically marked.)

Saad's methodology is firstly stylistical, then statistical. We choose to inverse the process. Indeed, we want the results of the text indexation to lead us to our textual hypothesis: is there a lexical change when there is a change of narrative voice?

We provide three results of our corpus indexation:

1. The 'palette' of Rebecca and its occurrence in the extract.
2. The 'palette' of the narrator and its occurrence in the extract.
3. The hypertextual reading verifies if the lexical changes follows a generic rule in the marks of the feminine gender (substantives, adjectives and determinants).

Our objective is to see whether the change of lexicon could be proven by the omnipresence of the feminine gender of the vocabulary used by the narrator. Therefore, is the polyphonic aspect of the narration an artifice used by Mandiargues to give an androgynous aspect to the narrator's account?

For this study, we use a lexical data processing software called TACT, created at the University of Toronto by John Bradley during the Eighties. The French interface was written by the professors Russell Wooldridge (University of Toronto) and Émilie Devriendt (École Normale Supérieure de Paris) and is available online on the French department's web site at the University of Toronto. We used MAKEBASE and USEBASE for the indexation process and TACTSTAT for the display of results. As the software itself is in English, we had to make some technical adjustments with respect to the non-equivalencies of the English and the French languages (for example accents indexation, which is an important part of the french language that can change the form of the '*vocable*', and therefore its grammatical category).

Our poster emphasizes the link between statistics and its theoretical field: stylistics.

Références

Delcroix, M. ; **Fernand H.** (sous la dir. de). 1990 [1987]. *Introduction aux études littéraires. Méthodes du texte*. Bruxelles, Duculot.

Giquel, B. (1999). *Stylistique littéraire et Informatique*. Arras, Artois Presses Université, coll. Cahiers scientifiques de l'Université d'Artois, n°8.

Muller, Ch. (1977). *Principes et méthodes de statistique lexicale*. Paris, Hachette Université, coll. Langue – Linguistique – Communication.

Pieyre de Mandiargues, A. (1963). *La Motocyclette*. Paris, Gallimard, coll. L'Imaginaire.

Saad, G. (1996). « La prose d'André Pieyre de Mandiargues : *ut pictura poesis* », dans *Des récits poétiques contemporains*, Sylviane Coyault (sous la dir. de), université Blaise Pascal, C.R.L.M.C., p.113-122.

Viprey, J.-M. (2002). *Analyses textuelles et hypertextuelles des Fleurs du mal*. Paris, Honoré Champion, coll. Lettres numériques.

Wooldridge, R. ; Devriendt, E. *TACT et TACTweb, Logiciels de recherche de données textuelles structurées*. Interface en français mise en ligne à l'URL http://www.chass.utoronto.ca/~wulfric/articles2/poitiers2001/

Managing a Short -Term Graduate Level Text Encoding Project

Caroline LEITCH

Department of English, University of Waterloo

This poster presentation will demonstrate not only the development of a digital scholarly edition of a mid-nineteenth-century social reform pamphlet, but also the successful management of a short-term encoding project at the graduate studies level.

The period between 1830 and 1860 saw a dramatic increase in the production of social reform texts in Britain. The question of how to alleviate the suffering of the labouring classes was explored in novels, poetry, pamphlets, and letters. While excellent collections of literary and non-literary social reform texts exist, what is lacking is a comprehensive research database of social reform writing that will make lesser-known non-literary texts available to scholars. I am currently creating a digital scholarly edition of one such text for my master's thesis. This electronic textual editing project is the first in a proposed series of editorial projects that will make non-literary social reform texts available in digital form.

My master's thesis involves both editorial practice and literary analysis. The editorial project will culminate in the creation of a digital scholarly edition of "Slaves of the Needle," a pamphlet written by Ralph Barnes Grindrod in 1844. "Slaves of the Needle" is representative of mid-nineteenth century British social reform rhetoric. Creation of the digital scholarly edition will involve encoding the pamphlet's structural features and using interpretive markup to identify the text's rhetorical patterns.

The second focus of my master's thesis is an analysis of the depiction of seamstresses in nineteenth-century British literature. The results of this analysis will be included in the digital edition of "Slaves of the Needle." Once completed, the digital scholarly edition will illustrate prevailing concerns about the labour conditions of female garment workers in mid-nineteenth century England. In the future, the digital edition of "Slaves of the Needle" will be part of a body of work concerning mid-nineteenth century social reform texts.

This project faces two challenges. The first challenge is posed by the text itself. The pamphlet includes a number of different types of text, such as letters, poetry, and eyewitness accounts, each with its own encoding difficulties. Before I begin creating the digital edition, I must research existing electronic textual editing projects and determine what methodology I will use. Because I will be using an interpretive markup scheme, I must decide which features I will identify and how I will classify rhetorical figures.

Creating a digital scholarly edition is an ambitious project. In this case, the project is made all the more challenging by the temporal limitations of a master's program. Estimated time from the project's inception to its completion is nine months. As opportunities for graduate-level digital humanities projects increase, students and advisors will be faced with the task of managing this kind of short-term encoding project. This poster presentation will show the digital scholarly edition in progress and discuss the project management process.

Generating Hypertext Views to Support Selective Reading

Eva Anna LENZ
Faculty of Cultural Studies
University of Dortmund, Germany
lenz@hytex.info

Angelika STORRER
Faculty of Cultural Studies
University of Dortmund, Germany
angelika.storrer@uni-dortmund.de

I. User Scenario and Goals

In many contexts, e.g. in interdisciplinary research, in scientific journalism and in specialised lexicography, readers search for information in a scientific domain in which they have previous but no expert knowledge. Their time is constrained, and they have to solve a very specific type of problem. In scenarios like these, users often read excursively and perceive only parts of longer documents. When these documents are sequentially organised, i.e. designed to be read from the beginning to the end, this selective reading may result in coherence problems. For example, a reader, jumping right in the middle of a sequential document, may not understand (or may misunderstand) a paragraph because he lacks the prerequisite knowledge given in the preceding text. The approach presented in our presentation generates hypertext views on sequential documents with the goal of avoiding coherence problems and making selective reading and browsing more efficient and more convenient than it would be possible with printmedia.

II. Strategies for the Generation of Hypertext Views

In contrast to other approaches to text-to-hypertext conversion, we generate hypertext views as additional layers while preserving the original sequence and content of the sequential documents. Thus, readers still have the option to perceive the documents in their original sequential form, provided they have the time to do so. The hypertext views mark an additional offer for those readers who only have the time for selective reading.

Our approach processes information coming from two levels:

1) On the document level the documents are annotated with regard to three annotation layers:

   - On the "document structure layer" we annotate structural units (such as chapters, paragraphs, footnotes, enumerated and unordered lists) using an annotation scheme derivated from DocBook.

   - On the "terms and definitions layer" we annotate occurrences of technical terms as well as text segments in which these terms are explicitly defined.

   - On the "cohesion layer" we annotate text-grammatical information of various types, e.g. co-reference, connectives, text-deictic expressions (cf. Holler et al. 2004).

While the annotation was performed manually in the first phase of the project, we currently investigate methods for automatic annotation (cf. Storrer & Wellinghoff 2006).

2) On the domain knowledge level we represent the semantics of technical terms occurring in the documents in a WordNet-style representation that we call "TermNet".

Using the annotations of the document structure layer, our hypertext views are generated using the segmentation principle "one-paragraph-is-one-node" as a starting point (which we refine in various ways). The rationale behind this is the expectation that paragraphs should be self-contained, and can thus be used as basic "building blocks" for a hypertext (Hammwöhner 1997). To support the selective reading of the resulting hypertext nodes, we enrich these nodes using two types of strategies:

1) Reconstructing cohesive closedness: Paragraphs in sequential documents often contain cohesion markers that point to information located external to the node – either in the preceding or in the subsequent text. Examples of such cohesive features are connective particles ("furthermore"), anaphoric

expressions ("he", "this issue") and text-deictic expressions ("the aforementioned specification"). In our hypertext views we try to reconstruct cohesive closedness by liberating these cohesive markers from their linkage to a specific reading path. For this purpose, we implemented four basic operations:

- *Anaphora resolution*: In the case of anaphoric expressions, its antecedents will pop up (cf. figure 1).

- *Linking*: Text-deictic expressions such as "the aforementioned specification" are transformed into links connected to the respective text segments.

- *Deletion*: Connectives, which first and foremost serve the creation of a fluent text (e.g. "yet"), are deleted.

- *Node expansion*: When a connective is directly bound to the preceding or subsequent text (e.g. "furthermore"), we provide the option to expand the current node and display the preceding or subsequent paragraph. With this option users may accumulate as much context as they need for properly understanding the node's content.

2) *Linking according to knowledge prerequisites*: With these types of strategies we offer additional information, which may be helpful for selective text comprehension. In our approach we concentrate on information related to the meaning of technical terms, because for our user scenario – the rapid search for information in a scientific domain – technical terms play a central role. In our hypertext views we offer two options to support the selective readers to better understand the terms and their underlying concepts:

- On the basis of the annotation layer "definitions and technical terms", we link the terms occurring in the documents to their definitions within the same document (cf. figure 2).

- On the basis of our TermNet, we generate glossary views, which show how a given term is linked to other terms and concepts of the domain (cf. figure 3). These glossary views also contain hyperlinks to text segments, in which the respective terms are explicitly defined. The glossary views are connected to all term occurrences in the documents; but the glossary can also be used as an additional stand-alone component.

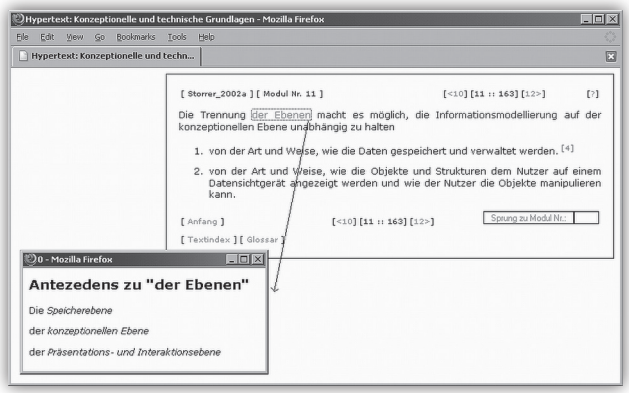


Figure 1: The NP anaphora "der Ebenen" (engl. the levels) is coreferent with each of the text segments "Die Speicherebene" (storage level), "der konzeptionellen Ebene" (conceptional level) and "der Präsentations- und Interaktionsebene" (presentational level) which occur in the preceding text. All three antecedents are displayed in a popup window.

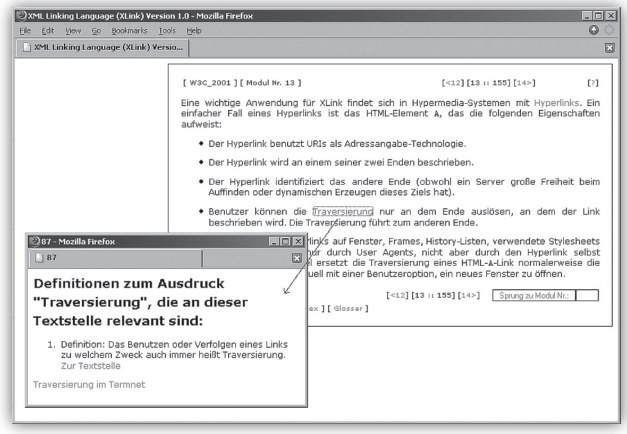


Figure 2: Popup showing a definition of a term.

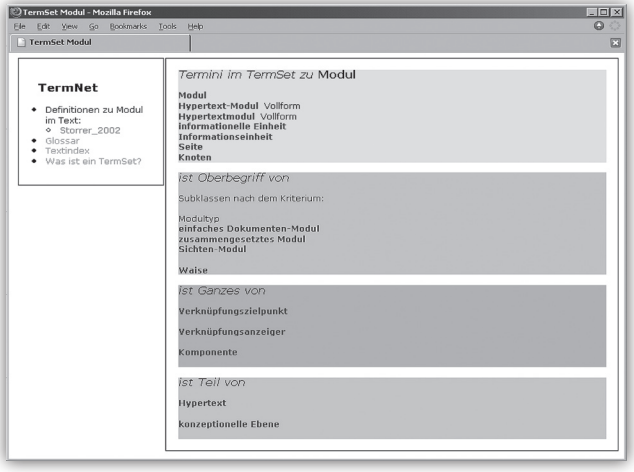


Figure 3: Glossary view of the term "modul", including a link to its definition (top left).

We implemented our strategies using a German corpus of 20 non-fictional documents (103.805 words) belonging to two specialised research domains, namely text technology and hypertext research.

III. Implementation Issues

All data – the TermNet as well as the annotated documents in our corpus – are represented and processed using XML technology.

TermNet is represented using the Topic Map Standard (Pepper & Moore 2001). To provide easy access to the data, we have developed an XSLT key library for XML Topic Maps (a key being comparable to an associative array in other programming languages). We use this library to perform simple consistency checks and to automatically infer additional TermNet relations, which may be of interest to the hypertext user browsing the glossary. In a subsequent step the TermNet, together with the newly inferred relations, is transformed into the hypertext glossary, again using the key library. Accordingly, the TermNet can be stored and maintained without redundant and error-prone information.

The documents of our corpus are annotated with respect to the three annotation layers described above: (1) the document structure layer, (2) the terms and definitions layer, (3) the cohesion layer. Following the approach developed by Witt et al. 2005, we store the three annotation layers in separate files. Thus, each layer can be annotated and maintained separately and can be validated against its corresponding document grammar (DTD or schema file). In a subsequent unification step, the different annotation layers of our corpus documents are merged. The resulting unified representation is the basis for another XSLT transformation, which automatically generates the hypertext views along the guidelines of our linking and segmentation strategies:

- Information from the "document structure layer" is used to generate an overall layout, to extract a table of contents and to perform text segmentation.
- On the basis of the "terms and definitions layer", we generate hyperlinks from technical terms to an ordered list of definitions for that term (the order being based on definition type and the position of the term relative to the definition).
- The "cohesion layer" is used for the reconstruction of cohesive closedness by means of one-to-one links, one-to-many links, anaphora resolution, deletion and node expansion.

All hypertext features are implemented using HTML and JavaScript as the target language of the XSLT transformation. This XSLT transformation is straightforward, but it has one disadvantage: each time when a linking or segmentation strategy is modified, or when the document grammar (of one of the three annotation layers) is changed, it is neccessary to adjust the (rather complex) programming code. In our project, where hypertextualisation strategies are to be tried out and tested, this turned out to be tedious (and sometimes error-prone) work.

In order to address this problem and facilitate the flexible testing and modification of our strategies, we designed the Hypertext Transformation Language (HTTL), a declarative language suitable for expressing rules of hypertextualisation (cf. Lenz in preparation). The language was designed with a hypertext expert in mind who writes a set of rules for linking and segmentation strategies; it offers general rules operating on abstract hypertext notions such as "hypertext nodes", "one-to-many links" or "node expansions". The expert may formulate segmentation and linking rules using HTTL, which is easier to learn than XSLT code, and allows for flexible experimenting with and refinement of hypertext conversion rules. HTTL also allows the user to use different rule sets for different situations. For example, a more coarse-grained segmentation strategy can be chosen for long text types.

Once stated, these rules are compiled into an executable XSLT program that performs the actual transformation of the annotated sequential documents into hypertext. Although some of the linking rules in HTTL may become rather complex (allowing for the generation of, e.g., one-to-many links with ordered link ends or node expansions), hypertextualisation rules in HTTL are much more concise – and thus more maintainable – than the generated XSLT code.

References

Hammwöhner, R. (1997). *Offene Hypertextsysteme*. Konstanz: Universitätsverlag Konstanz.

Holler, A., Maas, J.F. and Storrer, A. (2004). *Exploiting coreference annotations for text-to-hypertext conversion*. In: Proceedings of LREC, May 2004, Lisboa. 651-654.

Lenz, E. A. (in prep.). *Hypertext Transformation Language (HTTL)*. In: Dieter Metzing and Andreas Witt (eds.): Linguistic modeling of information and Markup Languages. Dordrecht: Springer. In preparation.

Pepper, S. and Moore, G. (2001). *XML Topic Maps (XTM) 1.0*. Topic-Maps.Org specification, March 2001. URL http://www.topicmaps.org/xtm/1.0/.

Storrer, A. and Wellinghoff, S. (2006). *Automated detection and annotation of term definitions in German text corpora*. Accepted for LREC 2006.

Witt, A., Goecke, D., Sasaki, F. and Lüngen, H. (2005). *Unification of XML Documents with Concurrent Markup*. Lit Linguist Computing, 20(1):103–116, 2005.

Fixing the Federalist: Correcting Results and Evaluating Editions for Automated Attribution

Shlomo LEVITAN
levishl@iit.edu

Shlomo ARGAMON
argamon@iit.edu

Linguistic Cognition Laboratory
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616, USA

Introduction

In the history of authorship attribution, the analysis of The Federalist Papers plays an important role. However, most previous non-traditional (stylistic) authorship studies have been flawed, mainly due to the use of improper editions, as documented by Rudman (2005). Our goal in this work was to perform a correct study by using a revised corpus of the Federalist papers based in large part on Rudman's critique. We used machine-learning techniques for analyzing the use of lexical features for authorship attribution of the papers. Another goal of our study was to explore how different corruptions of the corpus may affect the accuracy of the classification results, and the differences between them.

The Federalist Papers were written during the years 1787 and 1788 by Alexander Hamilton, John Jay and James Madison. These 85 propaganda tracts were intended to help to get the U.S. Constitution ratified, and were all published anonymously under the pseudonym "Publius". According to Avalon project (Yale Law School) Hamilton wrote 51 of the papers, Madison wrote 15, Jay wrote five, while three papers were written jointly by Hamilton and Madison, and 11 papers have disputed authorship – either Hamilton or Madison, although most evidence points to Madison as the author.

The Federalist Papers have presented a classic research problem for authorship attribution, and many studies on them have been done since Mosteller and Wallace's seminal attribution study (1964). Professor Joseph Rudman (2005) argues that in previous studies, textual flaws in the corpus were not properly addressed. To summarize Rudman's critique, problems in previous versions of the corpus included: Wrong letters and misspellings, due to letters typed by mistake or typos; inconsistency in inclusion of greetings and signatures in the papers; inclusion of foreign language words and quotations from other sources in the analysis; and inconsistent treatment of footnotes. Therefore, there is a doubt regarding the results of previous studies based on a flawed corpus.

The Corpus

We constructed a new corpus based on the one made available by the Avalon project, which is a collection of various papers that are related to United States history available on-line under the supervision of Yale Law School. The Avalon version is currently the most accurate on-line version of the Federalist Papers. To create a more accurate edition of the Papers, we compared the Avalon corpus to the edition of the Federalist book collection (*THE FEDERALIST: A COLLECTION OF ESS*AYS, 1788, Special Collection, Northwestern University) and corrected the flaws mentioned above. We used XML tags to group words by marking them specifically in the corpus in order to include or exclude them while processing the corpus. This option was used for comparison between the corrupt and corrected versions of the corpus or between the corpus with or without the groups. For example, footnotes, the remarks that the editor included in the corpus, were ignored by using XML tags while processing the corrected version. We addressed several problems in the corpus wrong letters and spelling in the papers, which were marked and fixed, according to the original source. Quotes and footnotes, were marked and fixed according to the original source in order to process the text with and without the marked data. Endings and openings that marked the text openings ("the People of the State of New York:" etc.) and the endings of the papers ("PUBLIUS"), were used in order to process the text with and without the marked data. These corrections address most of Rudman's criticism. In this work we disregarded punctuation because processing of the corpus was done automatically by extracting words, and this processing method doesn't take any punctuation issues into consideration.

Attribution Study

After we fixed the corpus, we counted various words to compute feature values. This operation included the extraction of frequent words and frequent collocations from the fixed corpus. We defined frequent words as the k most frequent words in the corpus where k is a parameter of the system. Frequent collocations (Argamon) and (Hoover) were defined as pairs of words occurring within a given threshold distance (window size) between them (for example, "for", "are" and "sure" appearing within 10 words of each other in a sentence, with a window size of 10). Given such a threshold, the most frequent such collocations were determined over the whole corpus. Given each particular feature set (frequent words, or collocations), the method was used to represent each document as a numerical vector, each of whose elements is the frequency of a particular feature of the text. For example, the word "the" was represented in a numerical vector as the number of times it occurs in the text divided by the number of words in the text. We then applied the SMO learning algorithm (Platt) with default parameters, which gives a model linearly weighting the various text features. SMO is a support vector machine algorithm; these have been applied successfully to a wide variety of text categorization problems (Joachims).

To analyze our results and to examine the accuracy of the classification we used 10-fold cross-validation, which is a common and reliable technique, to examine the generalization error of the model. 10-fold cross-validation divides the whole data set into 10 subsets of equal size; trains 10 times, each time leaving out one of the subsets from training, and then averages the results.

To find out the effect of corruption of the corpus, we performed 10-fold cross validation on the known Hamilton and Madison documents in both corrupt version and the corrected version of the corpus and compared the accuracy of results. We also examined the consistency attribution by building a model using all of the known training documents and classifying the 11 disputed papers, comparing the attributions derived from the corrected corpus as well as various corrupt versions of the corpus and various features sets. This experiment

allowed us to evaluate the effect of a corrupt corpus and to evaluate the benefit of creating a corrected corpus as suggested by Rudman (2005).

Results

The 10-fold cross validation experiment on the corrected corpus vs. the corrupt corpus produced results that allowed us to examine which corpus is better suited for further research. Our findings are described in *tables 1, 2 and 3* indicate that the experiments performed on the corrected corpus using different feature sets produced either slightly more accurate or similar results to the ones produced on the corrupt corpus. Our findings support the argument that the corrected corpus is a better source for further study of the federalist papers attribution problem.

We also addressed the authorship attribution problem of the 11 disputed papers, by building a model using the known papers as training documents and classifying the 11 disputed papers. Our results are summarized in *table 4* and show that the experiment conducted on both corpora, the corrected and the corrupt, produced the same results. The results clearly attribute the authorship of the disputed papers to Madison. Our findings are thus consistent with the universal accepted allocation that disputed papers were authored by Madison (Carey and McClellan).

Discussion

We have constructed a more accurate corpus of the Federalist papers, which was corrected and is now available for further research. Furthermore, the evaluation and analysis on both corpora were done by using modern machine learning methods that are completely automated. Our results show that using the corrected corpus for authorship attribution studies produces slightly better results than the corrupt corpus. Thus, our corrected corpus may be a better source for further study of the federalist papers attribution problem.

Moreover, this study provides additional support to the almost universally accepted allocation that Madison is the author of the disputed Federalist papers. We will be making our new corpus available for public use and we hope that it will become a useful tool for the research community in the future.

Conclusions

The corrected corpus of the Federalist papers that we have generated in this study was found to be a better source for further study of the Federalist papers attribution problem than the corrupt version. The experiments we conducted on the corrected and the corrupt corpus by using different feature sets provided results that support this argument. Furthermore, our study supports the universal opinion that Madison is the author of the disputed Federalist papers.

Acknowledgments

The authors would like to thank Joseph Rudman for his advice on all aspects of this project.

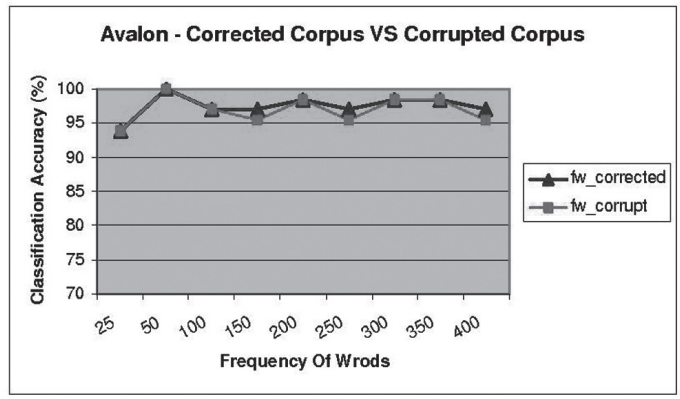


Table1. Classification Accuracy results for 25 – 400 most frequent words on Avalon corrected corpus VS. Avalon corrupt corpus.

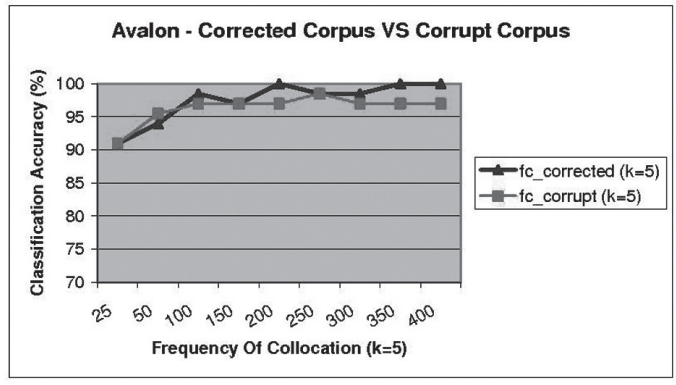


Table 2. Classification Accuracy for 25 - 400 most frequent collocation (window size 5) on Avalon corrected corpus VS. Avalon corrupt corpus.

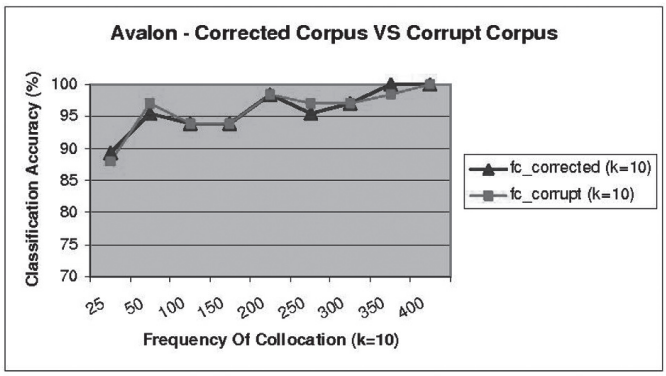


Table 3. Classification Accuracy for 25 - 400 most frequent collocation (window size 10) on Avalon corrected corpus VS. Avalon corrupt corpus.

Frequency Of Words	Papers Classified as Madison's Corrected Corpus	Papers Classified as Madison's Corrupt Corpus
25	10	10
50	10	10
100	10	10
150	11	11
200	11	11
250	11	11
300	11	11
350	11	11
400	11	11

Table 4. Disputed papers classification for most frequent words on Avalon corrected corpus vs. Avalon corrupt corpus.

References

Rudman, J. (2005). *The Non-Traditional Case for The Authorship of the Twelve Disputed "Federalist" Papers: A Monument Built on Sand*. Association for Computers and the Humanities and the Association for Literary and Linguistic Computing.

Rudman, J. (2005). *Unediting, De-editing, and Editing in Non-traditional Authorship Attribution Studies: With an Emphasis on the Canon Of Daniel Defoe*. The Papers of the Bibliographical Society of America, 99.1 pp 5-36

Mosteller, F. and **Wallace**, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, Mass.: Addison Wesley.

Avalon project Yale Law School (http://www.yale.edu/lawweb/avalon/federal/fed.htm)

*THE FEDERALIST: A COLLECTION OF ESA*AYS (As agreed upon by the federal convention September 17, 1787, in two volumes). Publisher: New-York: Printed and sold by J. and A, M'Lean 1788. The source was found in Library of Special Collections/ Northwestern University.

Argamon, S., **Levitan** S., (2005). *Measuring the Usefulness of Function Words for Authorship Attribution*. Association for Computers and the Humanities and the Association for Literary and Linguistic Computing.

Hoover, D.L. (2004). *Frequent collocations and authorial style*. Literary and Linguistic Computing 18.3 261-282

Goutte, C. (1997). *Note on free lunches and cross-validation*, Neural Computation, 9(6):1246-9.

Joachims, T. (1998). *Text categorization with Support Vector Machines: Learning with many relevant features*. In Machine Learning: ECML-98, Tenth European Conference on Machine Learning, pp. 137-142.

Platt, J. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, Microsoft Research Technical Report MSR-TR-98-14.

The federalist, the Gideon edition, Edited by George W. Carey and James McClellan.

A Context-Sensitive Machine-Aided Index Generator

Shelly LUKON
Patrick JUOLA

Dept. of Mathmatics and Computer Science, Duquesne University, USA.

Back-of-the-book indexing is the process of generating a list of relevant terms from a corpus and providing the user with the page references of these terms. It differs from web indexing in its ability to identify synonyms and subcategories. This indexing process has become somewhat automated through computer applications, most of which can at best generate a concordance, or list of keywords in the text. The difficulty lies in the ability to make intelligent decisions regarding which words or phrases to include. Human indexers primarily do this step. The biggest drawback is the time required to perform this task. It is estimated that every hundred pages of text takes one week to index manually (Chicago, 2003). The challenge, which the authors of this paper hope to have met, is to develop a software program that bridges the gap between computerized concordances and manual indexing, to provide a much more robust draft index, which a human indexer can refine in a fraction of the time. This application takes the ideas put forth in (Juola, 2005) and extends them to a working prototype system.

There are many different types of software available today for performing parts of the indexing process. Natural language parsers determine which words or phrases are subjects. Concordance-generation packages, such as AntConc, use a cluster analysis technique to study the relationships between words and their surrounding text, but still require the user to provide the specific words to search. Another application that uses cluster analysis, Grokker, aids in searches of text by creating visual maps of related terms. Many indexers use products like CINDEX, MACREX, and SKY Index, which function like databases, and can be integrated with Microsoft Access and dbase III. Each row of data consists of a heading, a subheading (if applicable), and the page number (locator). These programs take user-entered index information and provide an interface for grouping, censoring, sorting, and finally generating a clean index as output. No single application that is commercially available today, however, handles all of the processes described above. It is the intent of the computer-aided application described herein to do all of these things.

Our software application takes advantage of existing natural language parsers, and uses a factor analysis technique, based on the principles of Latent Semantic Analysis (Landauer, 1998 and Wiemer-Hastings, 2004) and Latent Semantic Indexing (Deerwester, 1990), to study the relationships among nouns / phrases and their surrounding text. This type of multivariate analysis uses matrix manipulation of the words and their locations in sentences, to create clusters of related words and contexts, and will assign probabilities to these clusters of words that can be translated into degrees of importance. A certain cutoff level of importance will be used to limit the size of the resulting list. The proximity of words to other words will also help to determine which words appear to be synonyms (terms that correlate highly in every dimension), and which words seem to be subcategories of broader terms (these terms, hyponyms, correlate highly in one or more dimensions.) In our application, LSA is used to generate the synonyms that will translate to cross-references and the hyponyms that will translate to subheadings of terms in the resulting index.

There are several resources currently available that define what makes a good index. One such resource is the NISO technical reference, "Guidelines for Indexes and Related Information Retrieval Devices" (Anderson, 1997). Anderson defines an index as "a systematic guide designed to indicate topics or features of documents in order to facilitate retrieval of documents or parts of documents." He further states that an index should include the following components: (a) terms representing the topics or features of documentary units; (b) a syntax for combining terms into headings in order to represent compound or complex topics; (c) cross-references among synonymous and other related terms; (d) a procedure for linking headings with particular documentary units; (e) a systematic ordering of headings (in displayed indexes). It is the objective of our application to successfully incorporate each of these components.

The task of evaluating our application's overall performance (i.e., how "close" the resulting indices are to

fulfilling the guidelines above) is a difficult one, which we attempt to answer by performing side-by-side comparisons of an index generated for a given corpus by our application and by a typical human indexer. We then compare the indexes generated by each, and create a matrix comparing (i) what we and the indexer both included in the index, (ii) what we included but the indexer did not, (iii) what the indexer included but we did not, and (iv) what neither of us included. We also note the relative times required for performing the task in each case. Our feeling is that if we can generate a similar index in a fraction of the time, then this application is a success.

It is our hope that this all-inclusive software, a "one-stop shop" for back-of-the-book indexing needs, will revolutionize the indexing industry by providing an easy to use, accurate means of generating an index in a drastically reduced amount of time.

References

Anthony, L. (2006). *Ant Conc 3.1.2 concordance generation software*, http://www.antlab.sci.waseda.ac.jp/

American Society of Indexers. (2006). http://www.asindexing.org/site/index.html.

Anderson, J. (1997). *NISO Technical Report 2: Guidelines for Indexes and Related Information Retrieval Devices*. 8. Bethesda: NISO Press.

Deerwester, S., et al. (1990). "Indexing by latent semantic analysis." *Journal of the American Society for Information Science, 41(6)*, 391-407. Wiley.

Groxis, Inc. *Grokker software*, http://www.groxis.com. San Francisco: Groxis, Inc.

Indexing Research. *CINDEX Indexing Software*, New York, NY, http://www.indexres.com. New York: Indexing Research.

Juola, P. (2005). "Towards an Automatic Index Generation Tool." *Proceedings of the 2005 Joint Annual Conference of the Association for Computing and the Humanities and the Association for Literary and Linguistic Computing (ACH/ALLC 2005).*

Jurafsky, D., Martin, J. (2000). "Word Sense Disambiguation and Information Retrieval." *Speech and Language Processing*, 631 – 666. NJ: Prentice Hall.

Landauer, T., et al. (1998). "An Introduction to Latent Semantic Analysis." *Discourse Processes*, *25*: 259 – 284. Mahwah, NJ: **Erlbaum Associates.**

Liu, H., MIT Media Lab. (2004). *MontyLingua: An end-to-end natural language processor with common sense*, http://web.media.mit.edu/~hugo/montylingua/.

Macrex Indexing Services. (2005). *Macrex Indexing Program*, http://www.macrex.com. Daly City, CA: MACREX.

O'Grady, W., et al. (2001). "Computational Linguistics." *Contemporary Linguistics 4th Edition*, 663 – 703. Boston: Bedford/St. Martin's.

Press, W., et al. (1992). "Singular Value Decomposition." *Numerical Recipes in C: The Art of Scientific Computing*, 59 – 70. Cambridge: Cambridge University Press.

SKY Software. SKY Index 6.0 Professional Edition, http://www.sky-software.com. Stephens City, VA: SKY Software.

Smith, L. (2002). "A Tutorial on Principal Components Analysis." www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf.

University of Chicago Press Staff. (2003). *The Chicago Manual of Style*, *15th Edition*, Chicago: University of Chicago Press.

Wiemer-Hastings, P. (2004). "Latent Semantic Analysis." *Encyclopedia of Language and Linguistics.* Oxford: Elsevier.

Synoptic Gospels Networked by Recurrent Markov Clustering

Maki MIYAKE

Department of Human System Science,Tokyo Institute of Technology

In this research, we represent the lexical co-occurrence information on the Synoptic Gospels under the form of a graph where the vertices correspond to the words or the concepts and the edges to the semantic paths. The aim of this study is to take advantage of the semantic network that is made for the Synoptic Gospels. This network is automatically constructed by a graph clustering method, a powerful technique to classify and reconnect the words in documents. To generate the network, we are now challenging to apply to the Gospels our original clustering algorithm (Jung, 2005), whose key idea derives from Markov Cluster Algorithm (MCL) that has been developed by van Dongen (2000).

Since the semantic network of the Synoptic Gospels is now articulated with lexical clusters that can be taken as "concepts" respectively, it might permit us to find the overview of the biblical world as well as to discuss the (genealogical) relationship among the Gospels, whose features will be projected onto some parts of the graph and traced at the same time by the flow of conceptual association.

Clustering Algorithm

For automatically drawing a concise semantic panorama of the world into a graph, we have recently proposed a new graph clustering algorithm, called Recurrent Markov Clustering (RMCL) method, which is one of the derivative methodologies of MCL.

The original MCL is based on random walks on a graph, and its model simulates flow by using two simple algebraic operations expansion and inflation on the stochastic transition matrix. This algorithm is applied in several research fields such as biology, linguistics and psychology. It is worth enumerating, for instance, Tribe-MCL for clustering proteins by Enright et al. (2002), Synonymy Network of Gfeller (2005) created with the addition of noise data, and Lexical Acquisition by Dorow et al. (2005) where some MCL clusters are merged for reconnecting concepts areas.

The new concept, RMCL uses the output of MCL as its input again. This improvement allows us to extract from the MCL results crucial information of clustered entities and their hierarchical relationships. RMCL is intended for a suitable control of the sizes of concept areas by the way of changing the granularity of the graph and the generality of the concepts.Articulating the recurrent process, we go back from the resulting converged state toward any on-going clustering step. This reversal procedure is the core part of RMCL to generate a virtual adjacency matrix of the hard cluster-nodes obtained through MCL process. This downsized matrix represents a simpler graph of the concepts built up by the similar words. We introduce here one of the recursive methodology, called Stepping-stone type algorithm, which consists of combining the particular clustering stages in progress of MCL with the final converged clustering stage.

RMCL process steps as follows:

Step1: MCL process with an adjacency matrix, where each node represents a word.

Step2: Reversal procedure to build a virtual adjacency matrix, where each node represents a cluster of MCL.

Step3: Repeated MCL process with the virtual adjacency matrix to compute a hard clustering for closing the cycle of a RMCL process.

Methodology

The following steps describe how to generate a network graph for the Synoptic Gospels by applying RMCL to the lexical co-occurrence data obtained from them:

1) Word Pairs Data obtained by a windowing method

Before using MCL process, it is necessary to make a list of word pairs that co-occur within a certain range of text. We practiced the windowing method to do this, which lead us to get a simplified representation of similarity level suitable for clustering. The technique is that the

window of a certain size slides over the sequences in a text to thoroughly extract fixed-sized grams (for example, tri-gram) of words (Vechthomova, 2003). The pairs were made afterwards by the combination of all the extracted words.

In the case of the Synoptic Gospels, the windowing method is applied to the Greek texts of the NT26th version by Nestle and Aland (1979), where a verse is set as document boundary. Various data are to be collected by changing window size from 2 (words) through 10. But we extract solely the word pairs which are simultaneously common to Mark, Luke and Matthew so as to focus on the common aspect of the Bible and show only the data of the window size 2 to manipulate 4930 word pairs which are composed of 769 word occurrences.

2) Dictionary-based Stemming

Since we are interested in the lexical-semantic information rather than the lexical-grammatical one, we took the stem form from words. The Stemming process was performed manually with BibleWorks Greek New Testament (BNM) Morphology. Additionally, 73 noise words were eliminated such as articles, prepositions, pronouns and conjunctions.

Finally we obtained a list of 1053 pairs with 468 word occurrences, and applied RMCL process to a 468*468 adjacency matrix.

3) RMCL Steps

At the step1, starting from the adjacency matrix of co-occurrence, the MCL process generated a nearly-idempotent matrix at the 13th cluster stage with 119 hard clusters. The reversal procedure computed a virtual adjacency matrix for each stepping cluster at the step2, and here we got 12 kinds of matrices. At the step3, we applied once again the MCL process to each virtual adjacency matrix obtained at the previous step. The reuse of the adjacency matrix generated with the 2nd cluster stage made the repeated MCL process flow until the 8th loop, where the number of the RMCL hard cluster turned out to be 65.

Results and Discussion

At the step 3, each intermediate cluster stage generates different adjacency matrix. As for selecting a particular one to be the most appropriate for interpretation, the variance of the RMCL cluster sizes can be considered as a criterion. The high variance means that the clusters are properly diversified to represent the multiple features of the text.

According to this criterion, we select the RMCL hard clusters of the 2nd cluster stage. Taking into account the result of MCL process, each hard cluster could be taken this time as a semantic or concept category.

If the symbol of "{ }" can be used for representing a component of words related with one another, there are some interesting components extracted, such as {"be pleased", "beloved"} or {"sleep", "die"}. The antonym category is also found as in {"forgiveness", "sin"} and {"dead(noun)", "live"}, the latter component representing the concept of "resurrection". The component of {split, curtain temple} makes us think of the presence of the idiom category, and in fact this one is referred to in Mk15:38. We have also found the topic category, in the components. For example, {new, fresh, old, wine, wine-skin} precisely means "new testament".

For the RMCL clusters, the largest component cluster included almost 30% of concept nodes. As the contents of this cluster can be described in terms of Jesus' redeem and passion, we can conclude from this that it represents as a whole the synoptic gospel genre itself.

Conclusion and Ongoing tasks

In conclusion, the application of RMCL to the Synoptic Gospels permitted us to create a compact semantic network in biblical lexicography, where the subject categories typical of the bible can be identified as concept clusters linked one another. Furthermore, it might be possible to use RMCL as a sort of ontology generator for the biblical studies, because we recognized that some of the main RMCL clusters contain a set of key words capable of producing taxonomic schemes or meta-data.

Obviously, the results of MCL clustering are more or less influenced by the initial selection of the co-occurring pairs, that is, the fixation of the breadth of the windowing frame. We are now working on the estimation of the appropriate values of parameters together with the morphological manipulations respecting various viewpoints such as Redaction criticism and Collocation (Syntax).

Our final goal is to accomplish the ontology of the Synoptic Gospels, which would be able to represent the fundamental categories of the Bible with the information on their mutual relationship.

Further information and more detail results will be found at the URL: http://home.a04.itscom.net/hilolani/sgn.htm.

References

Enright, A. J., Van Dongen, S.,and Ouzounis, C. A. (2002). *An efficient algorithm for large-scale detection of protein families*, Nucleic Acids, 30(7),1575-84.

BNM: BibleWorks LXX/OG Morphology and Lemma Database (2004). BibleWorks Greek New Testament.

Dorow, B. et al. (2005). *Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sence Discrimination*, MEANING-2005.

Gfeller, D., Chappelier, J.-C., De Los Rios, P. (2005). *Synonym Dictionary Improvement through Markov Clustering and Clustering Stability*, International Symposium on Applied Stochastic Models and Data Analysis, 106-113.

Jung, J., Miyake, M., Hatanaka, N., Akama, H. (2005). *For the Development of Composition Support System based on Semantic Network by Repeated Clustering*, IPSJ SIG-CE, No123, p99-105.

Nestle-Aland (1987). Novum Testmentum Graece 26th edition, German Bible Society Stuttgart.

Van Dongen, S. (2000). Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht.

Vechthomova, O., Roberston, S., Jones, S. (2003). *Query expansion with long-span collocates*, Information Retrieval, vol6, p251-273.

A Pilot Project to Assess the Suitability of the Voice-based Conferencing System Horizon Wimba for Use by Distance Education Students Registered in Second Language Courses at Athabasca University, Canada's Open University.

Audrey O'BRIEN
Kathy WILLIAMS
Centre for Language and Literature, Athabasca University

Corinne BOSSE
Educational Media Development, Athabasca University

Introduction

At Athabasca University, Canada's Open University situated in Alberta, approximately 1400 students register each year in second language courses (German, Spanish, English as a Second Language and French). In a distance education setting such as this a major challenge for both students and instructors is to find a satisfactory way of dealing with the oral component of language courses. At present students practise language skills by listening to the cassettes or CDs contained in their course materials and complete oral exercises and exams by telephone with their assigned tutor. Some phonetics exercises are also available in the digital reading room of the university library. The aim of the WIMBA pilot project was to ascertain whether the use of on-line voice communication might aid students to increase their oral and aural competence in second language learning. We also hoped to establish an on-line community of learners by providing students with the opportunity to communicate with each other and so compensate for the lack of interaction which is inevitable in a distance

education setting.

Implementation of the project

For the pilot project coordinators sent an e-mail to students to request volunteer participants. The participants were recruited from students registered in French and German language courses. Students were provided with a microphone.

A WIMBA Resource Website was created to orient the participants and to give them access to the system. During the orientation phase of the project, course coordinators introduced themselves to the students via the voice discussion board and asked students to reply by introducing themselves. Any technical malfunctions or questions from students regarding the operation of WIMBA were dealt with during this phase.

During the implementation phase coordinators put the course oral exercises on the WIMBA system and asked the students to practice and consult the coordinator for feedback until they felt ready to complete the actual oral exercise over the phone. Except for overseas students, students were still expected to complete the oral exercises over the phone in the usual manner. Overseas students had the opportunity to use voice e-mail to complete the oral exercises.

In addition to the course oral exercises one main discussion board was created for each language group to encourage peer interaction. The French and German course coordinators were active participants/moderators in these activities, although students were encouraged to use the voice board without instructor intervention.

At the end of the testing period a questionnaire was sent to students to elicit their opinion about the usefulness of the WIMBA system. The evaluation of the pilot project asked students, course coordinators and the WIMBA administrator to evaluate the project based on ease of use of WIMBA, increased oral communication, improved oral communication, and the success of building an on-line community.

Results

The student responses to the questionnaire were generally positive.

A high percentage of students (85%) agreed or strongly agreed that using the computer assisted language learning system helped improve their oral communication, 15% of the students were neutral that the medium helped their speaking skills. Likewise, the greater part of students (92%) agreed or strongly agreed that using the medium improved their listening skills and 8% of the students were neutral. All students reported listening to the voice boards more than once and 71.4% of students listened to the voice boards more than five times. However, even though a majority of participants agreed that WIMBA helped improve oral and aural communication, 60% of the student participants never used WIMBA to communicate with their peers.

Implications of these findings

The pilot project seems to suggest that we might have to take into account the self-directed and "silent learning community" of students who mentioned benefitting from listening to their peers' voices without necessarily responding. Even though students did not post to the boards containing oral exercises, the survey shows that they listened to these and used them to practise. This may imply that they listened to the introductions posted on the introductory voice board, and that they thus "got to know" their fellow classmates even though they did not actively engage in conversation with them.

At first glance the results from the pilot project seemed discouraging since we had hoped for much greater active involvement from the participants. However it may be that we need to rethink how researchers and the teaching community define and assess language learners' participation in a distance education context. Generally our students are self-directed learners, though some need more encouragement to achieve learner autonomy. The challenge for language instructors and course developers at Athabasca University is to help students improve their oral and aural competence whilst maintaining the flexibility of the distance education model. Students are working independently and will therefore not all be at the same point in a course at the same time, but we feel that it is important to offer students a forum which allows them to practise their language skills and discuss the course material. We would like to find a way of integrating WIMBA into our language courses and increasing students' awareness of the value of participating in the voice-boards without making this a compulsory part of any course. To this end we hope to continue experimenting

with the WIMBA system in order to find out which discussion subjects are most likely to get a response from students and to find ways of improving performance on oral exercises and exams.

Eventually WIMBA may be used in other areas such as literature and culture courses where voice discussion would enhance the learning experience of students.

The (In)Visibility of Digital Humanities Resources in Academic Contexts

Nikoleta PAPPA
Claire WARWICK
Melissa TERRAS
Paul HUNTINGTON

School of Library Archive and Information Studies, University College London

This poster aims to determine how easy is to find digital resources for humanities use, and whether links are made to such resources or portal sites (such as Humbul and AHDS) from library and humanities department websites. We have undertaken this task because one hypothesis about low levels of usage of digital resources by humanities scholars is that there is a lack of knowledge about appropriate resources. (Warwick, 2004) Thus if we determine that it is difficult for users to find information or resources, and that digital humanities projects are not promoted within academic contexts, then this would support such an explanation.

Context

This research is being undertaken as part of the LAIRAH project (Log Analysis of Internet Resources in the Arts and Humanities, (http://www.ucl.ac.uk/slais/LAIRAH/)), a collaboration between two research centres at University College London (UCL): the Centre for Information Behaviour and the Evaluation of Research (CIBER, (http://www.ucl.ac.uk/ciber)), and the newly created Cultural Informatics Research Centre for the Arts and humanities (CIRCAh (http://www.ucl.ac.uk/slais/circah/)). The aim of the LAIRAH survey is to investigate what influences the long-term sustainability and use of digital resources in the humanities through the analysis and evaluation of real-time use. This will provide comprehensive, quantitative, qualitative, and robust measures for evaluation of real-time use, utilising deep log analysis techniques on automatically recorded server

data, and undertaking user analysis of digital humanities projects. The results of this research should increase understanding of usage patterns of digital humanities resources; aid in the selection of projects for future funding, and enable us to develop evaluator measures for new projects.

Since 1998, over 300 projects have been funded by the Arts and Humanities Research Council in the UK to produce digital resources in the humanities, but while some of them have made an impact on humanities scholarship and are frequently used by scholars others have been relatively unsuccessful indicating financial and intellectual wastage. Limited research has been done into whether user centred factors determine usage (Warwick et al, 2005)

Through early CIRCAh group discussions we identified a limitation in the knowledge of how users access and locate digital resources for the humanities. Information seeking itself is a well researched area. As Dalton and Charnigo argue, in recent years the amount of literature about scholars and their information needs and information seeking behaviours has become a flood. Although useful recent work on humanities scholars has been done by Talja and Maula (2003), Greene (2000) and Ellis and Oldman, (2005) much of the literature tends to conflate information seeking and information needs in relation to humanities scholars. In the area of information needs, seminal work done by Stone (1982) and Watson Boone, (1994) showed that humanities users need a wide range of resources, in terms of their age and type. This is still true in a digital environment, where humanities users continue to need printed materials, or even manuscripts as well as electronic resources, which by their nature may imply a much greater age of materials than those used by scientists. (British Academy, 2005) Nevertheless, we are not aware of a study that has looked specifically at how or indeed whether humanities users seek digital resources. Thus this relatively small scale study should give rise to interesting findings, to be compared with the more general work of others.

Methods

Our study uses the methodologies of chaining, a term first coined by Ellis (1993), to describe the location of resources by information seeking by making links from references in text, and of browsing rather than key word searching. This is because these techniques are known to be widely used by humanities scholars (Bates, 2002). Humanities scholars do of course use keyword searches on Google. But to gain meaningful results from such a search scholars need to have enough knowledge to decide which terms to use. They may therefore prefer to chain or browse before they do so. Therefore the researcher, who is an experienced information searcher but does not have a background in digital humanities - thus eliminating the effect of knowledge transfer - will attempt to find digital resources for humanities scholar. The researcher plays the part of an interested but not particularly expert humanities scholar or graduate student researching a project or looking for information about a specific topic

Our sample takes 16 UK Russell Group universities, and will compare these to a sample of 16 UK post 1992 universities. The two types of university were chosen to represent a sample of the range of different universities in the UK higher education system. New universities are in general not as research intensive as older universities, and may teach subjects in different ways. Our sample was therefore chosen to avoid a concentration on one type of institution. The researcher searched for links to digital humanities resources and projects by using the university library websites and through websites for humanities departments or, if they are evident, scholars' web pages within these sites.

A total of eight History departments, eight English departments, eight Modern Languages departments and eight Film or Media studies departments were investigated. In each case four were from a new university and four from an old one. The universities were not duplicated (in other words we did not use both the English and History departments from UCL, but to chose those from different universities).

The researcher also used websites for the few research centres or courses on digital humanities that exist, as a starting point for her search. In the UK only 6 universities run such courses or host research centres. They are:

1. Centre for Computing in the Humanities, King's College London (U.K.)
2. Centre for Technology and the Arts, De Montfort University, Leicester (U.K)
3. Humanities Advanced Technology and Information

Institute, University of Glasgow (U.K.)

4. Pallas (Humanities Computing), Exeter University (U.K.)
5. School of Library, Archive and Information Studies, University College London (U.K)
6. Humanities Research Institute, Sheffield University, UK

These were chosen because scholars may choose to start searches from the pages of known centres of excellence in digital humanities research and teaching.

Results

Initial findings indicate that locating and using digital resources for the humanities is a difficult and time consuming task which may help to account for low levels of use. For example, from eight universities providing courses in film and media only two have digital resources links in their course/department web site. One has an easy accessible digital resources library catalogue, two have digital resources library catalogues but it is almost impossible to find them. The user has to be expert in information seeking in order to locate and access them (the researcher was able to find them only by accessing the web site map, something that a less experienced user would find it difficult to find). Finally three of the eight universities had no link at all in digital resources for film and media specifically and humanities in general. This represents only a quarter of our expected results, and we have chosen to present a poster because at the time of submission research is not yet complete. Thus we do not yet know how typical such results may be. However, the detailed results of further topics in this study will be analysed and presented at the conference to illustrate our findings, should the poster be accepted.

Conclusion

This study should prove illuminating, because of the lack of work that has been done on how humanities scholars find digital resources when beginning from web sites themselves. In the LAIRAH project, our overall aim is to discover the factors which make a resource usable, and one of these must be visibility. If users cannot discover that a resource exists, evidently they cannot use it. Our study therefore presets a small but important subset of results, which will inform our final findings about how resource creators and information professionals might make digital resources more usable to future researchers.

References

Bates, M. J.,(2002) *The cascade of interactions in the digital library interface*. Information Processing & Management. 38: 3 381 -400

British Academy (2005), *E-resources for Research in the Humanities and Social Sciences* - A British Academy Policy Review. section 3.5 Available from http://www.britac.ac.uk/reports/eresources/report/sect3.html#part5

Dalton, M. S, and Charnigo, L. (2004) *Historians and their information sources*. College & Research Libraries. 65 (5) 400-425.

Ellis, D. (1993) *Modelling the Information Seeking Patterns of Academic Researchers - a Grounded Theory Approach*. Library Quarterly 63 (4): 469-486

Ellis, D. and Oldman H. (2005) *The English literature researcher in the age of the Internet*. Journal of Information Science, 31 (1): 29-36.

Green, R., (2000) *Locating sources in humanities scholarship: The efficacy of following bibliographic references*. Library Quarterly. 70 (2): 201 -229

Stone, S. (1982) *'Humanities Scholars-Information needs and uses'* Journal of Documentation. 38 (4): 292-313

Talja, S. and Maula, J. (2003) *Reasons for the use and non-use of electronic journals and databases - A domain analytic study in four scholarly disciplines*. Journal of Documentation. 59 (6): 273-291

Warwick, C. (2004) *'Digital resources and Print Scholarship'* in Ray Siemens, Susan.Schreibman and John Unsworth (eds.) Blackwell Companion to digital humanities. Oxford, Blackwell. pp. 366-383.

Warwick, C., Blandford A. and Buchanan, G. (2005) *User Centred Interactive Search: A study of humanities users in a digital library environment*. Presented at

the Association for Computers and the Humanities-Association for Literary and Linguistic Computing, Conference 2005. University of Victoria, Canada, June 15-18.

Watson-Boone, R. (1994) *The Information Needs and Habits of Humanities Scholars*, Reference Quarterly, 34, 203-216

User Requirements for Humanities Digital Libraries

Jon RIMMER
Claire WARWICK
School of Library, Archive and Information Studies, University College London

Ann BLANDFORD
Jeremy GOW
UCL Interaction Centre, University College London

George BUCHANAN
Department of Computer Science, University of Swansea

Introduction

This proposal describes an initial set of end user studies conducted as part of the UCIS (User Centred Interactive Search with digital libraries[1]) project. Traditionally, digital library research has focused on improving system capabilities (such as the work reported by information retrieval literature), with little attention being paid to how information seeking behaviour develops over time and how this development can best be supported (e.g. Bates 1995). This project takes a three pronged approach in its research agenda: studying Humanities scholars' use of resources in context of the broader information task (such as writing), studying the development of expertise over the three year undergraduate degree program with Information Management students, and developing and testing novel interface features that support novice users and the development of expertise in the utilization of digital resources.

Context

With the advent of the World Wide Web there has been a dramatic increase in the digitization of information and artefacts. There is much discussion about how societies can benefit from this electronic delivery

of information as well as cautionary tales of problems associated with such a rapid take-up (Lesk 2004). The Scientific community has embraced the technologies facilitated by the World Wide Web, indeed creating and nurturing them. Humanities researchers have, in general, not been so quick to weave these resources into their research repertoire, with the exception of generic tools such as Google and online library catalogues and bibliographic tools. Reasons for this could include a lack of comfort and confidence with information technology, a reliance on colleagues and networking events as a source of information, reliance on their own personal collections, and a slower, more serendipitous way of searching and formulating their research ideas (see Barrett 2005 for an overview of these positions). Humanities researchers reportedly tend to find chaining, (or following up links from footnotes) to be a more useful activity than searching (Green 2000). Despite the hypertextual nature of the web, this activity is seldom well supported in online environments (Bates 2002).

Over these past few months the largest organisations that provide WWW search engine services have been competing to create impressive digital libraries that will give the widest possible audience access to very large bundles of resources (e.g. Mathes 2005 & BBC 2005). Many of the resources being donated or used by these projects will be of significant value to Humanities scholars. Our research is immediately relevant to such projects, since it offers a solid foundation of knowledge about the working practices, experiences and attitudes of their intended audiences.

Methodology

We are therefore studying humanities users, and their interactions with information, in both physical and virtual environments. An important facet of our work is that users are studied in as naturalistic a context as possible, to gain a fuller understanding of the nature of their information work. Data is being gathered using interviews, observations, electronic logging and diary studies of use of digital and traditional library materials. Three focused research agendas are being tackled, looking at Humanities scholars, postgraduate researchers and first year undergraduates. Data collection and analysis are ongoing; we currently have data from 25 interviews and observations, the majority being from English literature and history specialists, and expect the final paper to be based on detailed analysis of data from approximately 40 participants.

Findings and Discussion

Using broad grained discourse analytical techniques proposed by Potter and Wetherell (1987) a list of themes that have emerged from the transcribed interviews will be presented. Offered here is a brief overview of such themes:

Insights into the positive and negative aspects of the Humanities 'research experience'

Detailed descriptions of their research activities revealed the "Sherlock Holmes" nature of their work; how it develops across the use of many sources and how the 'mystery' is investigated by 'chasing up leads'. Additionally, the depths of engagement experienced during interaction with the actual source materials were described. So for example, hunting down a rare 16th century book in a second hand shop and slowly leafing through it over the weekend was described as a highly pleasurable, personal experience. This poses a significant design challenge: How can digital resources best support the work of the research 'Sleuth' and how can the experience of doing so be enhanced to facilitate engagement whilst interacting with technology?

The Physical and the Electronic (Real and Virtual)

Different experiences in a variety of physical libraries were discussed, and how these research experiences differed to the use of electronic resources was also explored. We shall be addressing how some of the qualities of the physical browsing activity can be best supported by electronic resources. This is being done by developing, prototyping and testing interfaces that offer additional information to the user in a variety of ways, such as statistics on article use, related material, and similar search pathways through the data.

Space, place and people

The importance of, and problems of, places (libraries, auction houses, book fairs), spaces (e.g. working in particular libraries) and the relationships with other people were also revealing. These findings can be set against electronic resources to see how well they support or hinder these relationships. Do these technologies need to consider ways of incorporating additional communication tools to support research communities?

How resources are assessed

The criteria scholars used to evaluate resources were often implicit. These interviews revealed issues of accuracy and ease of use for both physical and electronic resources. Our prototyped interfaces are exploring ways of expressing, for example, how results are ranked and how the user can interact with the system in order to present the data according to their own preferences.

Embracing technology

Participants discussed how different sorts of technology fitted into their research practices over the last 25 years, including first use of email, and more recently the Web and electronic resources.

Problems with technologies old and new

Critiques were offered of microfiche, microfilm, CD Roms as well as library catalogues and Internet search engines. By understanding barriers to previous technological take up in general, improved techniques can be developed to promote these resources to the Humanities research community.

The modern researcher and a need for training

The impact of Internet search engines on the quality of students' work was also discussed, and how there is now a greater need for the researcher to undertake more formal training to acquire the necessary skills to discriminate between information and their sources. Discussion of electronic resource use amongst a wide set of humanities scholars is informing how such training would best be structured and prioritised.

Discussion

This research is enabling us to develop, test and deliver open source working systems that support the development of expertise in information seeking with digital libraries. System development is based on the Greenstone software (www.nzdl.org), allowing us to focus on questions of interaction rather than developing our own technical infrastructure. Our research will provide a detailed account of the development of expertise in information seeking with digital libraries, and allow the development of models of information seeking behaviour with these resources, contextualised within an understanding of the broader information tasks (researching and writing) of Humanities academics.

These initial studies have been useful for us in the development of scenarios and personas that form the basis of the project's user requirements. These have proved to be highly useful to design teams in general (Jordan 2000) and are particularly useful in illustrating the user, and the context within which they work, to the technical designers on this project. The initial engagement with humanities scholars has gathered rich information that is shaping our initial designs. These subject matter experts will continue to participate in the project by evaluating our prototypes as part of an iterative design process.

Acknowledgement

This research is part of the User Centred Interactive Search project which is funded by the EPRSC grant number GR/S84798.

Footnotes

[1] http://www.uclic.ucl.ac.uk/annb/DLUsability/UCIS.html

References

Barrett, A. (2005) The information seeking habits of graduate student researchers in the humanities. *The Journal of Academic Librarianship*. 31(4), 324-331.

Bates, M. J. (1995) The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions. *College & Research Libraries*. 57, 514-523.

Bates, M. J. (2002) The cascade of interactions in the digital interface. *Information Processing & Management*. 38(3), 381-400.

BBC News (2005) Microsoft Scans British Library http://news.bbc.co.uk/1/technology/4402442.stm (visited 4th November 2005)

Green, R. (2000) Locating sources in Humanities

scholarship: The efficacy of following bibliographic references. *Library Quarterly*. 70(2), 201-229

Jordan, P. W. (2000) *Designing Pleasurable Products: An Introduction to the new Human Factors*. Taylor & Francis Books; U.K.

Lesk, M. (2004*) Understanding Digital Libraries*. 2nd Edition. San Francisco, CA: Morgan Kaufmaann.

Mathes, A. (2005) Preserving Public Domain Books. http://googleblog.blogspot.com/2005/11/preserving-public-domain-books.html

Potter, J. & Wetherell, M. (1987*) Discourse and social psychology: Beyond attitudes and behaviour*. London: Sage.

Introducing the Pattern-Finder

Stephanie SMOLINSKY
Humanities Department,
New York City Technical College, CUNY

Constantine SOKOLOFF
Independent Scholar & Programmer

We would like to present the Pattern-Finder, a new computational tool for analyzing texts with specific application to the close reading of poetry. It was designed and created by Smolinsky & Sokoloff, and can be found at www. patternfinder.net. Our program was invented to examine whether (and if so, how) the phonetic/phonological structure of a poem may contribute to its meaning and emotional power—to take a close look at how the 'music' of poetry actually works. Two Pattern-Finder studies have already been completed (on Keats' *Bright Star* sonnet, and on the Wallace Stevens lyrics *Bantams in Pine-Woods* and *Fabliau of Florida*). These, we claim, uncover aspects of poetic technique inaccessible to the naked ear; displays from these studies will form part of our presentation. A third study, on Hopkins' Terrible Sonnet *No Worst, There Is None*, is being proposed as a paper for this conference. Below, we describe briefly the theoretical background to our invention; then we describe how the Pattern-Finder is used on a text, and the type of sound patterning it is designed to bring to light.

Our approach comes from combining two sources, one literary and one linguistic. The first is the traditional critical close-reading focus on sound repetitions, on phenomena like alliteration, assonance, rhyme and meter (for the closest predecessor to our approach, see Burke, 1973). The second derives from the study of a topic in linguistics known as *phonetic symbolism*. This is an ongoing area of research, which started in the U.S. over 70 years ago with Edward Sapir's (1929) experimental *Investigation in phonetic symbolism*. Following Sapir, many researchers have continued investigating whether isolated speech sounds are not in themselves experienced as carrying some intrinsic meaning. There is growing evidence that they can indeed trigger attribution of qualities, such as smallness or largeness, brightness or darkness.

(See, among many others, Chastaing (1958, 1962) Fischer-Jørgensen (1978) Fonagy (1979) Magnuse (1999) Newman (1933) Smolinsky (2001) and Tanz (1971)).

In one part of Sapir's experiment, subjects were given two made-up words for "table," *mal* and *mil* (/mal/ and /miyl/ in the SAMPA (Wells, 1997) transcription system we are using). The subjects then had to decide which table was 'bigger' and which was 'smaller.' Since the vowels in these words were varied while the surrounding frame of consonants were kept the same, in effect, it was the perceived 'size' of the vowels which was being tested. Sapir, testing a range of vowels, found there was significant consistency of response, and Newman (1933) continuing Sapir's experiment four years later, found enough consistency in subjects' responses to go as far as setting up scales for vowels and consonants, showing brightest to darkest and largest to smallest.

Certain regularities in both Sapir's and Newman's findings suggest that the way such attributions of qualities to speech sounds may work is by *features*: "Phonetic propert[ies] used to classify sounds," (Ladefoged, 1982, p 38). For example, one such feature of consonants is *voicing* (whether they are made with or without vibration of the vocal cords, as in the voiceless /t/ and /s/, which have as their voiced counterparts /d/ and /z/). To illustrate, there is a pronunciation rule in English that when a noun ends in a voiceless consonant, its plural suffix will also be voiceless: *cat* /kAt/ pluralized is *cats* /kAts/. But if the noun ends in a voiced consonant, for example *dog* /dQg/, then its plural will also be voiced: *dogs* /dQgz/. This rule also has an important extra clause: if, as in *midge* /mIj/ or *roach* /rowC/, the noun ends in a voiced or voiceless consonant which is *coronal* (i.e., one made by raising the blade of the tongue: (Chomsky and Halle, 1968) and *sibilant* (characterized by marked loudness and a sustained high-frequency noise (Ladefoged, op cit.), then the plural suffix is realized as a whole syllable: *midges* /mIjIz/ or *roaches* /rowCiz/.

For our purposes, what is relevant about this kind of pronunciation rule, operating below the conscious level, is that if native English-speakers are given a series of made-up nouns, such as *drobe* /drowb/, *bloph* /blQf/ or *nutch* /nVC/, and asked to pluralize them aloud, they can, without any apparent effort, produce the correct plurals for all three unknown words. It seems that they automatically scan each noun's last sound for the presence of, respectively, voiced, voiceless and sibilant-plus-coronal features, and assign the suffixes accordingly. Thus, it can be strongly argued that features like *voiced* or *coronal* must be available in our subconscious for obeying pronunciation rules. If so, then they should also be available for symbolizing use, as in Sapir's "large" and "small" *mal* and *mil* examples. We, as readers, might be affected by such symbolizing material in a poem (for example, a sense of spaciousness or constriction) without quite knowing what in its music has triggered our response, any more than we 'know' why the plural of *mirage* is *mirages* with its extra syllable. The exploration of this hypothesis—that feature-patterning is the driving force in the 'music' of poetry—was the idea behind the creation of the Pattern-Finder.

The process of Pattern-Finder analysis is as follows: Firstly, we put the text into a (broad) phonetic transcription; this is an essential step, since a consistent relationship between sounds and their written representations is needed—one hardly provided by normal English orthography. We have chosen the SAMPA phonetic transcription alphabet (see Wells, op.cit.) because it uses the same characters as a normal keyboard: once the reader has the key to the system, it is immediately available for use. Prosodic features (three degrees of stress, two pause lengths) are also transcribed with keyboard characters. The text is then input to the Pattern-Finder with a search command for a feature or features; next, the program analyzes it for the feature(s) required, and finally the analyzed version of the text is output with highlighted displays of the distribution of the feature(s), together with the frequency count(s).

Most of the Pattern-Finder's features are the conventional ones found in any phonetics textbook, such as *high* for a vowel, or *bilabial* for a consonant. But a number of extra features were added, wider-ranging ones (such as *upper*, which includes both high and mid-high vowels, or *labial*, which includes any consonant made with the lips) and more narrowly-focused ones, such as *breathy* and *sibilant*, which distinguish between higher-pitched and lower-pitched sounds involving frication. In designing our range of features, the intention was to imitate as much as possible of the range of unconscious phonetic/phonological distinctions in the mind of the English-speaking poet and reader. In this way, we hoped to capture more of the sound-patterning, broad or narrow, which we hypothesize underlies the music of a poem.

The Pattern-Finder offers an extensive range of analyses, single or combinatory, segmental and/or prosodic. Users may analyze the text for up to four features in one search. They may request "one-highlight" displays, i.e., of a single set of speech sounds identified either by one feature (e.g., all the *high* vowels) or by two features (e.g., all the *high* and *back* vowels). Or they may request two-highlight "versus" dsiplays, i.e., of two sets of speech sounds identified by either one feature (all the *high* versus all the *low* vowels), or by two features (all the *high* and *back* versus all the *low* and *back* vowels). The availability of prosodic features means that users may also search, e.g., for all the *stressed* and *high* vowels in a text, or *stressed* and *high* vowels versus *pre-pause* and *voiceless* consonants.

The resulting highlighted displays of speech sounds, picked out by one or two features, allow the user to track their distribution throughout the text. In this way, one is able to match *feature-patterning* (for example, whether this sound appears in clusters at one or two places in the text, or more sparsely but at regularly-spaced intervals, or is almost, or totally, absent) to *meaning* (such as contrasts of atmosphere or subject, correspondences between apparently disparate images, locus of climax, and so on).

Naturally, we are most eager for the chance to share this program with our peers. (Besides English, we can also offer clear explanations in French, Russian, Spanish, Catalan, and Hindi.) We feel that we have invented a kind of X-ray machine for poetry—or, indeed, for longer texts—and we are keen to invite all who love literature to join with us in using this new instrument for investigating its sound-structures.

References

Phonetics and Phonological Features

Brosnahan, L F. and B. Malmberg (1970) *Introduction to phonetics,* Cambridge, UK: Cambridge University Press

Chomsky, N. and M. Halle (1968) *The sound-pattern of English*. New York: Harper & Row

Gimson, A. C. (1989) *Introduction to the pronunciation of English*. (4th edition) London: Edwin Arnold

Hyman, L. M. 1957 *Phonology: theory and analysis*. New York: Holt, Rinehart & Winston

Jakobson, R., G. Fant and M. Halle (1951) *Preliminaries to speech analysis: the distinctive features and their correlates*. Cambridge MA: MIT Press

——— and **M. Halle** (1971) *Fundamentals of language*. The Hague: Mouton

Ladefoged, P. (1962) *Elements of acoustic phonetics*. Chicago and London: University of Chicago Press

——— (1982) *A course in phonetics.* (2nd edition) New York: Harcourt, Brace, Jovanovich Inc.

Peterson, G. E and **H. L. Barney** (1952) Control methods used in a study of the identification of vowels, *Journal of the Acoustical Society of America* v 24 #2, pp 175-184

——— and **I. Lehiste** (1960) Duration of syllable nucleii in English. *Journal of the Acoustical Society of America* v 32, pp 693-703

Pickett, J. M. (1980) *The sounds of speech communication.* Baltimore, MA: University Park Press

Phonetic Symbolism and Literary Criticism

Burke, K. (1973) "On Musicality in Verse" in *The Philosophy of Literary Form.* Berkley, CA: University of California Press

Chastaing, M. (1958) Le symbolisme des voyelles, signification d'i, *Journal de Psychologie* v 55, pp 403-423 & 461-481

——— (1962) La brillance des voyelles, *Archivum Linguisticum* v 14, pp 1-13

Fisher-Jørgensen, El. (1978) On the universal character of phonetic symbolism with special reference to vowels, *Studia Linguistica* v 32, pp 80-90

Fónagy, I. (1979) *La métaphore en phonétique*. Ottawa: Marcel Didier

——— (1983) *La vive voix*. Paris: Payot

Grammont, M.(1946) *Traité de phonétique*. Paris: Delagrave

Jakobson, R. (1978) *Language in literature*. Cambridge, MA: Belknap Press of Harvard University

Jespersen, O. (1933) The symbolic value of the vowel i in *Linguistica: Selected writings of Otto Jespersen*. Copenhagen: Levin & Munksgard pp 283-303

Magnus, M. (1999) *Gods of the Word*. Kirksville MO: Thomas Jefferson University Press

Marchand, H. (1959) Phonetic symbolism in English word-formation, *Indogermanische Forschungen* v 64, pp 146-68, 256-277

Newman, S. S. (1933) Further experiments in phonetic symbolism, *American Journal of Psychology* v 45, pp 53-75

Ohala, J. J. (1994) The frequency code underlies the sound-symbolic use of voice-pitch, in *Sound Symbolism* (eds Hinton, L., J. Nichols and J. J. Ohala) Cambridge, UK: Cambridge University Press

Sapir, E. (1927) Language as a form of human behavior, *The English Journal* v 16, pp 421-433

——— (1929) A study in phonetic symbolism, *Journal of Experimental Psychology* v 12, 225-239

Smolinsky, Stephanie (2001) *Brilliance, energy and size in vowels: a cross-linguistic study of phonetic symbolism*. Unpublished Ph.D. dissertation, CUNY Graduate Center, NYC.

Tanz, C. (1971) Sound symbolism in words relating to proximity and distance, *Language and Speech* v 17, pp 87-94

Transcription

SAMPA www.phon.ucl.ac.uk/home/sampa

Wells, J. C., (1997) 'SAMPA computer readable phonetic alphabet'. In Gibbon, D., R. Moore and R. Winski, (eds.) *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B.

Creating and Using a Digital Version of Giovanni Villani's "*Nuova Cronica*"

Matthew SNEIDER
History Department,
University of Massachusetts-Dartmouth

Rala DIAKITE
Humanities Department,
Fitchburg State University

On a 1300 pilgrimage to Rome the Florentine merchant Giovanni Villani was inspired by the crumbling glories of the ancient metropolis. It occurred to him that his own city, the daughter and creature of Rome, was on the rise and had need of an historian to do the patriotic duty of a Virgil, a Sallust, or a Livy. Despite reservations about his ability to measure up to such eminent men, without, that is, the aid and grace of God, Villani resolved to begin a record of the "deeds and beginnings" of the Florentines and matters in the wider world. He would undertake this work "with reverence for God and Saint John, and in praise of our city of Florence". The result, his immense *Nuova Cronica*, traced the working of God's providence in the vicissitudes of Florence, Italy, and Europe; it glorified Florence as a worthy inheritor of the mantle of Rome; it revealed the righteous judgment of God against the sinful; it sought to provide a stock of moral and political *exempla* to future generations of Florentines.

We are engaged in an ongoing project to make this tremendously important chronicle more widely available to students and scholars: we are creating a fully encoded and annotated online version of the chronicle which will be accompanied by its first complete English translation.

This project is supported by an N.E.H. grant and is part of Brown University's Virtual Humanities Lab. The V.H.L. provides texts and databases which are encoded, annotated, and contextualized. These include the *Decameron Web*, the *Pico Project*, and the *Catasto/Tratte*. The long term aim is to create a kind of virtual Florence where students and scholars can access searchable

versions of fundamental texts placed in rich historical context. We hope that the streets and plazas of our electronic Florence will serve the same social and intellectual function as those of the original: we want the V.H.L. to be a forum for exchange and collaboration in learning and scholarship.

Professor Diakite and I began our work in the final book, which runs from 1342 through 1347. These were significant years for the history of Florence, Italy and Europe: the abortive lordship of Walter of Brienne, the disarray of Angevin power in Southern Italy, the stunningly successful campaigns of Edward III, and the first news of the Black Death. Villani shifts masterfully between the narrow stage of his native city and the events of the wider world; his treatment is nuanced and detailed. Our choice to "begin at the end" of the chronicle was dictated by the brilliance and the usefulness of this final book.

Our first task was to develop an encoding scheme. After much discussion about the potential scholarly and pedagogical uses of the electronic text, we crafted a scheme which takes in textual structure and textual elements, citations and sources, dates, place-names, personages, and recurrent themes.

After we had fully encoded the final book of the chronicle, we began producing annotations linking significant words and passages to ancillary material. Particularly important episodes will be linked to contextual essays and bibliographies; proper names will be linked to biographies and bibliographies; place names will be linked to descriptions and maps; references to sources will be linked, where possible, to the original texts. Perhaps most excitingly we plan to create links to alternate descriptions of the same events in other chronicles. The resulting web of medieval voices, Italian and European, will provide a complex and nuanced view of the major events of the 14th century.

The fully encoded and annotated *Nuova Cronica* will be a powerful tool. Students and scholars will have Italian and English versions of the chronicle, richly commented and woven into a web of complementary voices. They will be able to conduct complex and fruitful searches: a student of 14th century Italian politics, for example, could easily find every occasion where Villani moralizes against the tyrannous acts of lords while a student of 14th century economic history could find every occasion where Villani records a price. The most important dimension of our project, however, is the contribution we hope *users* will make to its development. We intend to "open the door" to scholars who are interested in working with the chronicle; they will be invited to add links to material and to create annotations. We also intend to create a forum for scholarly discussion and electronic publication. We hope, in other words, that our digital text will be the center of a collective *lavoro* and *lettura*.

Our poster will trace the challenges involved in encoding and annotating a work as complex as the *Nuova Cronica* and will provide a preliminary balance of our accomplishments. Its focus, however, will be the relationship between scholarly and pedagogical uses of the electronic text. This was a matter of some concern as we developed our project: how could we make our electronic text both relevant for scholars and useful for students? We will describe how we sought to match our encoding scheme to two very different sets of users. We will explore ways to integrate the chronicle, along with the other components of the V.H.L., into courses on Italian literature and history. We will present sample projects which would allow students to explore our text and to exploit its encoding and annotation. We will propose ways to involve students in the work of scholars through the posting and discussion of papers in online seminars and through the use of chat rooms.

The Buccaneers of America: A Multilingual Comparative Electronic Edition

Cynthia SPEER

Electronic Text Center,
University of Virginia Library

In 1678, the Amsterdam printer Jan ten Hoorn published *De americaensche Zee-Roovers*, a historical (yet understandably sensational) account of the pirates who preyed upon the Spanish treasure fleets and Caribbean colonies of the seventeenth century. Although somewhat obscure in origin, the text was later able to be attributed to Alexandre Olivier Exquemelin, a Huguenot ship's surgeon to the buccaneers who witnessed most of the events he describes. The book proved suitably tantalizing to readers to ensure its rapid translation into German (1679), Spanish (1681), English (1684), and ultimately French (1686). While making the text (known as *The Buccaneers of America* in English) available to readers in new countries and cultures, the translators of these editions also accomplished a second, somewhat more duplicitous, task: adding new material of their own which transformed the texts completely to suit the purposes of the societies for which they were printed. Subsequent editions of each different language version have thus only continued to corrupt the original text, further complicating the history of textual transmission of the original but ultimately elucidating the socio-political and ideological climates out of which they arose.

Participants in the Buccaneers of America project are attempting to prepare an interlinked "meta-edition" of the original Dutch text and its early translations into German, Spanish, English and French. The Dutch, English and French texts have been digitally photographed and tagged in basic TEI at the University of Virginia Library; German and Spanish editions from the Library of Congress have been photographed and will be TEI-encoded. One of the primary goals of the project being to aid scholars in their comparisons of the texts across different language versions, several means of interlinking the texts -- linguistically, structurally and thematically -- will be attempted. First, a selection of keywords from the texts will be encoded with cross-references to period and modern translations occurring in other parts of the editions: therefore searching and/or browsing for the key title term "bucaniers" in English will also retrieve results for "aventuriers" in French, "piratas" in Spanish, "See-Raeubers" in German, and "Zee-Roovers" in Dutch, plus the usual twentieth-century terms used in published titles. In addition, further structural and thematic comparisons between the editions will be made by the use of descriptive text collations. The web interface of the collations will feature passages of text encapsulated in brief summaries, categorized by narrative and thematic divisions, and presented in the order they occur in the text. Corresponding passages that "track" between the two editions will be linked to represent their association to the reader. Readers will also note that for some passages in one edition, there is no corresponding passage in another edition, as well as remarking that even when passages occur in both comparison editions, they may have been reordered. Links from each of the text divisions in the collation will take readers off to the corresponding section of the edition itself, or to desired paratextual material.

Although such keyword indexing as mentioned above is able to be at least partially automated, it still requires a certain amount of manual manipulation of the data (thus being somewhat labor-intensive), and remains at best an incomplete solution to the key issue of linguistic transparency across the editions, pointing to a key concern of project participants: accessibility and utility of the editions to interested scholars everywhere, especially in the Caribbean itself, with the smallest barriers to use possible. Related to this is the larger issue of cultural heritage and its "appropriation," however unintentional, by outside cultures and agencies involved in the editions' organization and execution.

Travelers in the Middle East Archive (TIMEA): Integrating Texts and Images in DSpace with GIS and Teaching Resources

Lisa SPIRO
Electronic Resources Center, Rice University

Marie WISE
TIMEA project, Rice University USA.

Like other digital archives, the Travelers in the Middle East Archive (TIMEA) acts as a repository for digital texts and images, in this case works documenting travel to the Middle East between the 18th and early 20th centuries. In TIMEA (http://timea.rice.edu), sponsored by the Institute of Museum and Library Services and Rice University's Computer and Information Technology Institute, one can find digitized versions of historic stereocards and postcards depicting such sites as the pyramids and the sphinx, along with TEI-encoded texts such as travel guides, travel narratives, and scholarly works.

Yet offering access to unique sources is only one of the project's goals. TIMEA, which is based at Rice University, also aims to provide valuable resources for teaching; improve information literacy and research skills; offer a model for building learning communities that use electronic resources; and develop innovative mechanisms for using GIS tools in cultural heritage projects. Thus TIMEA demonstrates the geographical nature of its focus by creating dynamic GIS maps that combine geospatial data such as elevation and water with layers displaying different historical maps and travel and trade routes. In addition, TIMEA addresses the critical need to cultivate information fluency and research skills and build learning communities. As studies have found, students lack essential skills in finding and using information: "University libraries have outstanding information resources available to their student populations… and they have powerful tools for accessing these materials… but many college students are either unaware of these resources or they do not know how to use them" (Quarton 120). By creating research and teaching guides, TIMEA develops fundamental methods among students at the same time that it gives access to a particular body of material. These guides are presented in Connexions (http://cnx.rice.edu), an open, collaboratively-authored repository of electronic course materials.

To deliver the texts and images, TIMEA uses DSpace, an open-source digital repository system. The choice of DSpace was driven by several factors: it is open-source, supports the Open Archives Initiative (OAI), and provides a long term archiving solution that will ensure access to the TIMEA digital assets well into the future. Moreover, Rice University's Digital Library Initiative recently adopted DSpace, making a commitment to support and develop the system. Initially DSpace was designed for research materials and scholarly papers, so it currently has several limitations for digitization projects: support for viewing XML documents in a user-friendly way is not yet available, there is limited support for structural metadata, and the interface design cannot easily be customized for each community within an institution's installation of DSpace. However, institutions such as Texas A & M and the University of Rochester are beginning to experiment with using DSpace for archives of digitized materials. TIMEA and Rice are contributing to these efforts by addressing the problems of XML support and interface design. In collaboration with Texas A & M, Rice is working on providing a customizable user interface for each DSpace community that is driven by Cocoon and XML, which will allow TIMEA to have its own unique look and feel. Likewise, Rice is developing support for XML publishing in DSpace. We are awaiting METS support for structural metadata, which is currently being developed and will be included in a future DSpace release.

By bringing together multimedia resources such as XML texts, digital images, and GIS maps with teaching and research modules, TIMEA faces a crucial challenge: integration, both technical and intellectual. In so doing, TIMEA is building on the work done by complex digital projects such as the Valley of the Shadow and Perseus. TIMEA content resides in three separate systems: texts and images in DSpace; GIS maps in ESRI's ArcIMS map server; and research and teaching modules in Connexions. By using these systems, TIMEA can leverage the unique functionality offered by each. Since TIMEA

is an ongoing project, its technical team continues to explore the best means for integration, such as a web services solution that leverages the availability of data in each system in XML. In large part, success depends upon active collaboration among the various contributors, including the GIS team, digital library systems developer, project managers, Connexions staff, and module authors. To provide interlinking among the texts, images, and Connexions modules, the project team is taking advantage of the permanent URIs for digital objects generated by DSpace and the rich linking capabilities for research modules in Connexions. In order to connect GIS maps with texts, the GIS support specialist has authored a program that automatically drops in links to map locations in the XML files based on place names. So that users can seamlessly navigate TIMEA's various elements, the digital library systems developer is working on the aforementioned project to provide a customized web interface in DSpace.

Even as the TIMEA team works through the technical issues of integrating a complex archive, it also faces questions about how to realize the project's scholarly and educational goals and serve its user community of teachers, students, researchers and museum professionals. How can the research and teaching modules be used to augment rather than overdetermine students' understanding of TIMEA materials? How can TIMEA provide links among the various components that lead to new understanding rather than overwhelm the end user? These questions are being addressed through collaboration among the project team members, active consultation with scholars and teachers, and user testing.

As a whole, we hope that TIMEA's components will come together to support the project objectives in a way that highlights the scope and depth of available resources. As a contribution to computing in the humanities, TIMEA is an example of how diverse resources can be integrated to enable more sophisticated means of conducting scholarly inquiry.

References

Henry, Geneva, Baraniuk, Rich, and Christopher Kelty, *"The Connexions Project: Promoting Open Sharing of Knowledge for Education,"* Syllabus2003 Conference, July 2003. http://www.syllabus.com/summer2003/pdf/T03.pdf

Perseus Digital Library. Accessed 2005-11-14. http://www.perseus.tufts.edu/.

Quarton, Barbara, *"Research Skills and the New Undergraduate."* Journal of Instructional Psychology. 30.2 (June 2003): 120-124.

Smith, MacKenzie, et al. *"DSpace: An Open Source Dynamic Digital Repository."* DLib 9.1 (January 2003). http://www.dlib.org/dlib/january03/smith/01smith.html

Texas A & M. TXSpace. Accessed 2005-11-14. http://txspace.tamu.edu/

University of Rochester. UR Research. Accessed 2005-11-14. https://urresearch.rochester.edu/index.jsp

Valley of the Shadow. Accessed 2005-11-14. http://valley.vcdh.virginia.edu/

Humanities Computing and the Geographical Imagination: The Mark Twain's Mississippi Project

Drew E. VANDECREEK

Northern Illinois University Libraries

The Mark Twain's Mississippi Project World Wide Web site (http://dig.lib.niu.edu/twain) presents users with a digital library featuring humanities materials shedding light upon the historical milieu in which Samuel Clemens grew to maturity, and which he remembered and imagined as Mark Twain in a series of celebrated works based in the Mississippi River Valley of the mid-nineteenth century. The site will present Mark Twain's Mississippi works (*The Adventures Tom Sawyer*, *The Adventures of Huckleberry Finn*, and *Life on the Mississippi*) online in a fully searchable digital format, along with other contemporary authors' descriptions, accounts, and definitions of that region. Project staff members have gathered and digitized over 100 primary source texts, including travel accounts, immigrants' guides, gazetteers, and reminiscences, from participating libraries, and are at work presenting them on the project web site. They have also gathered nearly one thousand images from these texts, as well as participating institutions' collections of visual materials, and are mounting these materials in a parallel database. Project staff have identified and gathered mid-nineteenth century sheet music totaling some twenty songs describing and mythologizing the Mississippi River, its valley, and its culture. Musicians have recorded versions of these songs, which will be featured in database of sound recordings. Finally, the project World Wide Web site presents spatial data via Geographic Information Systems technology, including geographical features and data sets depicting the changing demographic, economic, and political contours of the region in this period. Using these online tools, project users may compare Mark Twain's accounts of the Mississippi Valley of the nineteenth century with those produced by other observers, thereby exploring and analyzing significant themes in American literature and history.

Digital Humanities Quarterly

John WALSH
Michelle DALMAU

Digital Library Program, Indiana University

The proposed poster will highlight Digital Humanities Quarterly (DHQ), a new open-access, peer-reviewed, digital journal covering all aspects of digital media in the humanities. Published by the Alliance of Digital Humanities Organizations (ADHO), DHQ is also a community experiment in journal publication, with a commitment to:

- experimenting with publication formats and the rhetoric of digital authoring
- co-publishing articles with Literary and Linguistic Computing (a well-established print digital humanities journal) in ways that straddle the print/digital divide
- using open standards to deliver journal content
- developing translation services and multilingual reviewing in keeping with the strongly international character of ADHO

DHQ will publish a wide range of peer-reviewed materials, including:

- Scholarly articles
- Editorials and provocative opinion pieces
- Experiments in interactive media
- Reviews of books, web sites, new media art installations, digital humanities systems and tools
- A blog with guest commentators

The poster will be developed by DHQ's technical editor, with contributions from other editors and staff. In addition to providing general information about the journal, the poster will provide details about the technical development and infrastructure behind DHQ, including the TEI-based DQH authoring schema, the open source content management system that manages and delivers DHQ content, and the combinations of

XML, XSLT, CSS, and other Web technologies that drive the DHQ interface.

It is hoped that DHQ will become an important resource for the digital humanities community. The poster will serve both to promote an important new venue for scholarship and research in digital humanities and to provide valuable information about the technological development of a complex and evolving online resource.

The Virtual Mesoamerican Archive: Exploring Expansion Possibilities, Automated Harvesting, and Migration to MySQL

Stephanie WOOD

Wired Humanities Project,
University of Oregon

The Virtual Mesoamerican Archive (VMA) is work in progress. It is intended to be an extensive, academic, portal website. It aims to codify, integrate, and provide Internet links to the vast and internationally dispersed collections of digitized cultural and historical materials of early Mesoamerica (1800 BCE to 1800 CE) as provided by archives, libraries, museums, private collections, and individual scholars.

The potential audience for this reference material includes scholars, graduate and undergraduate students, secondary and elementary students and teachers, and a broad cross-section of the general public in many nations interested in Mexican and Central American history, whether because of their own heritage and/or a fascination with the cultural richness offered by the ancestors of these global neighbors (ancestors such as the Olmecs, Aztecs, Mixtecs, Zapotecs, Mayas, and others.).

Begun in 2002 by Stephanie Wood and Judith Musick at the Wired Humanities Project, University of Oregon, as a "proof of concept" effort, and currently only accessible in preview form, the VMA will allow users online access to thousands of records in five inter-related databases. We currently have over 5,500 records in these databases. The contents are selected by our subject experts, then gleaned manually from electronic and print sources, or solicited directly from institutions and individuals, with data entry performed by closely supervised undergraduate students. We are sorting this material into multiple categories, with extensive information and hyperlinks to:

- Repositories – museums, galleries, archives, and

libraries – that hold Mesoamerican cultural heritage materials.

- Collections held by each repository, including collection inventories when available.
- Digitized facsimiles of significant Mesoamerican cultural materials, with added keywords.
- Online articles and websites authored by scholars, with added keywords.
- Contact information, education details, publication records, and career histories for scholars who devote their time to Mesoamerican studies.

We have a number of tasks that remain, which we would like to explore with our peers at this ACH poster session.

- We aim to augment and expand content, not just manually, as we have been doing, but through the provision of online submission forms that would feed into holding tanks for the contents to be approved prior to being merged with existing databases.
- We also wish to explore the possibilities of automated data harvesting protocols, always with permission from targeted institutions. We can envision a use for this with sites that have hundreds or thousands of items of cultural heritage materials already available on line, such as the Justin Kerr Precolumbian Portfolio of 1,618 images, or the Peabody Museum online collections, with 10,109 Maya pieces, among others.
- We plan to improve and expand our search functionality, with better access to the materials we have accumulated. This will include embellishing our advanced search mechanisms and preparing for the browsing of preset bodies of data, anticipating users' interests.
- This past summer we undertook a test migration of our database contents and their presentation on the web from FilemakerPro to MySQL. This was successful, and we hope to fully realize this transition when the databases are more complete, to ensure accessibility and longevity for the project. MySQL and PHP compatibility may also facilitate automated data harvesting and mergers.
- We also have yet to encode the texts we have either authored ourselves, such as the biographical sketches of Mesoamericanist scholars, or texts we have harvested and inserted into our databases, such as from museum websites (always with their permission). We plan to use TEI Lite for this, since these texts are fairly short and simple. Online essays and websites, on the other hand, will not be encoded because we do not actually serve these, only brief descriptions and links to them.
- We have already begun establishing a professional advisory board, but we need to expand it. The board's responsibilities will include the formalization of editorial standards and procedures as well as the facilitating of inter-institutional partnerships to sustain the VMA.

The URL for the VMA preview:

http://whp.uoregon.edu/vma_preview/

username = aza

password = PERSIA

Ontology for a Formal Description of Literary Characters

Amélie ZÖLLNER-WEBER
amelie.zoellner-weber@uni-bielefeld.de

Andreas WITT
andreas.witt@uni-bielefeld.de

Bielefeld University, Text Technology

Introduction

A story plot would not work without the actions of the literary characters included. Actually, their actions and reactions do essentially form the plot. Furthermore, these characters are often provided with special behaviours or features enriching the story. Thus, within literature studies, the analysis of characters is an important task for analysing and interpreting literature. But the focus in recent approaches is often very restricted so that only some parts of literary characters are shown (Note [1]) or they demonstrate only a theoretical classification and do only partially provide evidence for the analysis on literature. Other approaches deal only with special characters so that a generalisation for all characters is often not possible. (Note [2]) Representing and describing a character in most of its facets is often not reached in a (nearly) complete way.

We will present an application of a formal ontology for literary characters. The concepts, which are included in the ontology, combine several theories for literary characters. The goal is to derive a model of the structure of literary characters. For this task, it is required to establish a flexible and open system so that newly found categories or aspects can be included without restructuring the ontology. Additionally, it should be possible to integrate the representation of many characters.

The definition of 'ontology' originates from philosophy and was transferred to the field of the 'Artificial Intelligence' (AI). In general, the philosophical theories of ontology are used to explain the existence of things in the world.(Note: [3]) When the terminus 'ontology' was introduced in the field of AI the definition of ontology underwent a semantic change. In AI ontology is defined as the modelling of concepts of the real world in computer systems. A definition is given by Gruber: "An ontology is a formal, explicit specification of a shared conceptualisation." (Note: [4]) The main goal of an ontology is to represent information in a structured way. Thus, an ontology comprises hierarchically structured concepts of a part of the 'world', called 'domain'. Additionally, there are relations between these concepts. Usually, an ontology is built up by 'classes' or rather 'concepts', 'properties' (also 'slots' or 'roles'), and restrictions. By adding so called 'instances' individual objects of a class can be realised. The representation of a formal ontology is expressed in a machine-readable format. A formal ontology is represented in form of a logical formalism. This has the advantage that the given information can be used as the base for inferring new information.

Since the 90s, ontologies representing information have become a focussed research field. They are used in many disciplines especially in linguistics or in life sciences. (Note: [5]) Adapting the concept of an ontology in the field of humanities computing is rather new. To our knowledge applying a formal ontology to the representation of literary characters has not been realised so far.

Modelling of the structure of literary characters

This approach combines several theories of literary characters to develop the base of the ontology.(Note: [6]) This results in a definition of literary characters as mental information structures of the person who analyses them.(Note: [7]) The information structure consists of a bundle of properties and features.(Note: [8]) It is not always possible to observe all aspects but it is important to take most of them as background for the analyses of literary characters.

In our approach, classes, properties, instances, and relations of the ontology together form a complex formal description of the structure of a character. A literary character is represented by several instances. This differs

from other ontologies because there, an individual, e.g. a special protein in the Gene Ontology, is represented as one instance. But the outlined ontology represents a feature or an information of a literary character as one instance. Otherwise, it is not possible to include all facets of a character.

Up to now, a first prototype of an ontology was realised. A rough set of classes and properties was developed. This set was modelled in the ontology editor 'Protégé' (Note: [9]). An extract of the ontology in this editor is shown in figure 1. Protégé allows the representation of a formal ontology in the W3C-Standard 'Web Ontology Language' (OWL).(Note: [10])

For applying the ontology, a corpus of selected characters is constantly integrated into the system. The corpus consists of selected 'devil' characters in literature covering the motive "Doctor Faustus". The group of devil characters are interesting because they show properties of human beings as well as properties of supernormal characters. The text corpus comprises an intersection of several literary epochs like 'Sturm and Drang' or the period of 'Biedermeier' and different kinds of text (e.g. tragedy, early prose, prose). A large part of the text corpus is already XML-annotated. The annotation is based on TEI P5 (Note: [11]), but user defined modifications of the tag set are included as well. An aim is to cross-reference from the ontology to parts in a text showing a special event of the character and vice versa.

After an evaluation phase the ontology system will be adapted to the needs of a potential user. Later, the ontology will be open for the community, so that interested users can describe literary characters and can do a search for them. Therefore, the inclusion of new categories for the ontology is allowed. Providing the ontology to the community enables the development of extensions as well as the inclusion of further characters. We expect that such an ontology can be a good resource for the comparative literature studies, especially since inferences across several characters can be drawn automatically.

Our approach is focussed on a community which is interested in literary characters and formal ontologies. Accessing the ontology has to be realised in an intuitive way so that a user does not need to have much knowledge about the applied technologies.

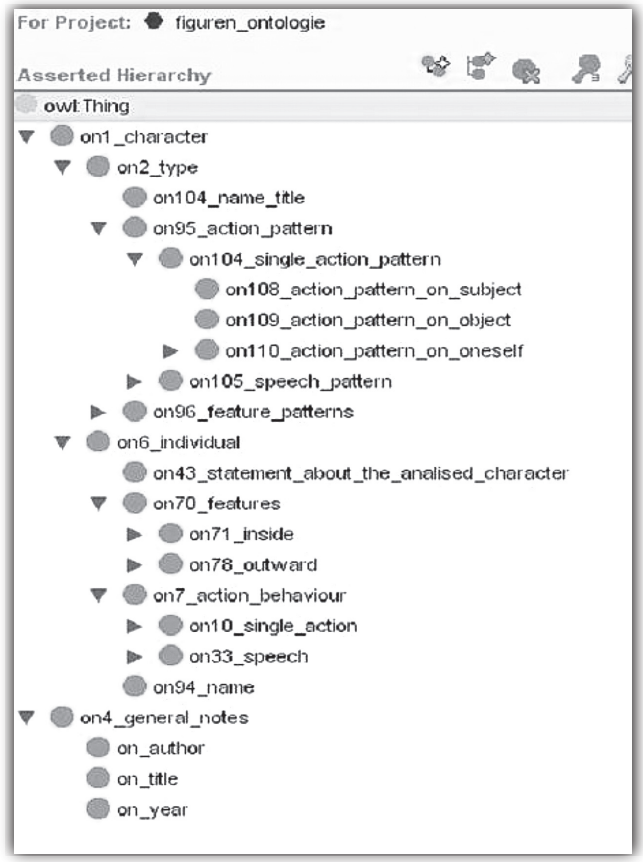


Fig. 1: Part of the ontology of literary characters in the Protégé editor.

Summary

The presented approach proposes a way to model the information about literary characters in a formal way. Making use of this concept it is not intended that the ontology contains a consensus of analysing and describing characters. In this respect our ontology differs from ontologies, from the "hard sciences", where an ontology is regarded as a "shared conceptualisation" of a field. By applying the technique of a formal ontology a flexible and extensible system can be realised. This offers new and individual perspectives on literary characters. Because the ontology is modelled with the rather new web technology standard OWL, a prerequisite for an open access is satisfied.

In the near future the functions for describing and searching characters will be improved. It is also planned to integrate more literary characters.

Notes:

[1] For example, Propp or Greimas analyse literary characters only as far as their functions for the plot are concerned (see Propp 1975, Greimas 1983).

[2] Wahl Armstrong 1979, Propp 1975.

[3] See also Puppe et al. 2000.

[4] Gruber 1993, p.199.

[5] An example of an application in natural science is the Gene Ontology (GO, http://www.geneontology.org). In linguistics many applications of semantic webs like WordNet (http://wordnet.princeton.edu), GermaNet which are very similar to an ontology are developed.

[6] Some categories of the approaches of Fotis Jannidis, Werner Kummer, Jurij M. Lotman, and Göran Nieragden are adopted (see Jannidis 2004, Kummer 1975, Lotman 1972, Nieragden 1995).

[7] See Jannidis, 2004, p.185.

[8] This definition relates to Lotman's approach (Lotman 1972).

[9] See http://protege.stanford.edu

[10] See http://www.w3.org/2004/OWL/

[11] See http://www.tei-c.org/P5/

References

Greimas, A.J. (1983). *Les Actants, les Acteurs et les Figures*. In: Algirdas Julien Greimas: Du sens II: Essais sémiotiques. Paris: Editions du Seuil. pp.46-66.

Gruber, T. R. (1993). *A translation approach to portable ontology specifications*. Knowledge Acquisition, 5. pp.199-220.

Jannidis, F. (2004). *Figur und Person - Beitrag zur historischen Narratologie*. Berlin: Walter de Gruyter.

Kummer, W. (1975). *Grundlagen der Texttheorie - Zur handlungstheoretischen Begründung einer materialistischen Sprachwissenschaft*. Hamburg: Rowohlt Taschenbuch Verlag.

Lotman, J.M. (1972). *Die Struktur literarischer Texte. Chapter: Der Begriff der Figur; Von der Spezifik der künstlerischen Welt, Figur und Charakter*. München: Wilhelm Fink Verlag.

Nieragden, G. (1995). *Figurendarstellung im Roman: Eine narratologische Systematik am Beispiel von David Lodges Changing Places und Ian McEwans The Child in Time*. Trier: Wissenschaftlicher Verlag Trier.

Propp, V. (1975). *Morphologie des Märchens*. Karl Eimermacher (Eds.). Frankfurt am Main: Suhrkamp.

Puppe, F. et al. (2000). *Knowledge Engineering*. In: Günther Görz et al.(Eds.): Handbuch der Künstlichen Intelligenz. München: Oldenbourg Verlag 2000. pp.599 - 641.

Wahl Armstrong, M. (1979). *Rolle und Charakter – Studien zur Menschendarstellung im Nibelungenlied*. Göppingen: Kümmerle Verlag.

Index of Presenters