SYNOPSIS

Applying Computational Linguistics and Language Models: From Descriptive Linguistics to Text Mining and Psycholinguistics

Gerold Schneider

Department of English
Institute of Computational Linguistics
University of Zurich
gschneid@ifi.uzh.ch

April 2014

Contents

Introduction

This synopsis presents the application of computational linguistic tools and approaches which were developed by the author for Descriptive Linguistics, Text Mining, and Psycholinguistics. It also describes how the computational linguistic tools, which are originally based on linguistic insights and assumptions, lead to new and detailed linguistic insights if applied to different research areas, and can in turn again improve the computational tools. The computational tools are based on models of language, predicting part-of-speech tags or syntactic attachment. These models, which were originally designed for the practical purpose of solving a computational linguistics task, can increasingly be used as models of human language processing.

A large-scale syntactic parser is the core linguistic tool that I am going to use. I further also employ its preprocessing tools, part-of-speech taggers and chunkers, and approaches learning from the data, so-called data-driven approaches. Parsing (from Latin *pars, -tis*) refers to the (automatic) assignment of syntactic structure to sentences. An example of a syntactic analysis in a Dependency Grammar representation is given in Figure 1. For example, the main verb *fear* is correlated to *experts* with a subject relation, and to a subordinate clause *the advertisements could actually make the problem worse*. The most important syntactic labels of the dependency representation used by the parser is summarised in Table 1. An introduction to Dependency Grammar (Tesnière, 1959) and a detailed discussion of this parser, Pro3Gres, including its prepocessing steps and evaluation, are given in Schneider (2008). The currently largest corpus collection that I have used comprises over a thousand million of words.

I only give a very brief overview of some of the parser's important aspects here. The workflow of the parser is illustrated in Figure 2. Such a modular architecture allows one to update components or replace them with improved tools. In the workflow, the raw text to be parsed is first part-of-speech tagged, then base noun phrases and verb groups (so-called chunks) are recognised, and the base forms of words (lemmata) are reported. Currently we use the C&C tagger, the morpha lemmatizer and the LT-TTT2 chunker (Grover, 2008). We have also used pipelines with different taggers, for example Tree-tagger (Schmid, 1994) or Ratnaparkhi's MaxEnt tagger (Ratnaparkhi, 1996), and different chunkers, namely LTPOS (Mikheev, 1997) and Carafe (Burger and Bayer, 2005). The actual parsing takes place between the heads of the chunks. The integration of chunking and parsing fits Dependency systems particularly well, as Abney (1996) point outs. Chunks largely corresponds to Tesnière's concept of *nucleus*. Figure 2 also illustrates that the permissible syntactic structures are licensed by a hand-written *competence* grammar, and filtered, ranked and disambiguated by statistical *performance* data learnt from the Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993). As competence grammars massively overgenerate, continuous filtering (pruning) is required.

The use of syntactic parsing opens up a wide range of possibilities. In the **first chapter**, I summarise my applications of syntactic parsing, its preprocessing tools, and other computational linguistic approaches for the benefit of Descriptive Linguistics. I describe collocations, language variation, alternations, and language change. I

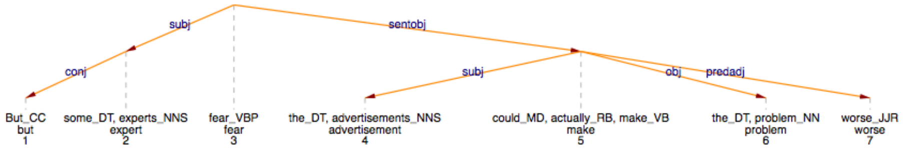


Figure 1: Syntactic Analysis in a Dependency Grammar representation of the sentence *But some experts fear the advertisements could actually make the problem worse*

RELATION (governor–dependent)	LABEL	EXAMPLE
verb–subject	*subj*	*he sleeps*
verb–direct object	*obj*	*sees it*
verb–second object	*obj*2	*gave (her) kisses*
verb–adjunct	*adj*	*ate yesterday*
verb–subord. clause	*sentobj*	*saw (they) came*
verb–pred. adjective	*predadj*	*is ready*
verb–prep. phrase	*pobj*	*slept in bed*
noun–prep. phrase	*modpp*	*draft of paper*
noun–participle	*modpart*	*report written*
verb–complementizer	*compl*	*to eat apples*
noun–preposition	*prep*	*to the house*

Table 1: Frequent syntactic labels of the parser Pro3Gres

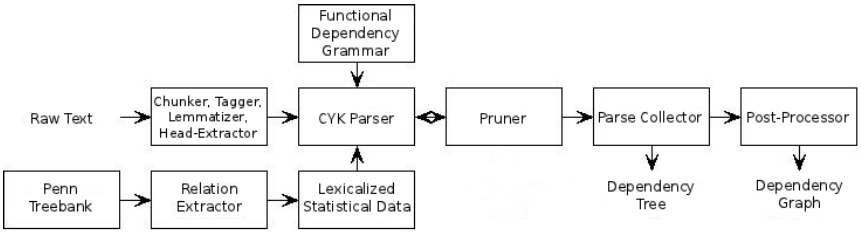


Figure 2: Pro3Gres flowchart

will also describe the obvious advantage of an automatic approach: the sheer amount of data that can be processed, and the consistency, which can lead to the data-driven detection of new patterns. I also focus on the obvious disadvantage of using an automatic tool: that there is always a certain level of errors, which entails that evaluations are essential.

In the **second chapter** I describe the application of the same tools for Biomedical Text Mining. I evaluate the performance of our approach and summarise insights from a linguistic perspective, leaving more technical aspects to the side.

In the **third chapter**, I argue that a syntactic parser, in particular my approach which draws a clear division between *competence* and *performance*, can be used as a model to explore formulaic and creative language use, starting with Sinclair's (1991) distinction between idiom principle and syntax principle, and ending with the suggestion to use the parser as a psycholinguistic model.

Throughout this study, I address the research question of how Descriptive Linguistics, Text Mining, and Psycholinguistics can profit from the use of computational linguistic methods. During my research, this overarching question unfolded into an array containing the following research questions. Some of these questions were central at the beginning, others arose from intermediate results.

- Can automatically parsed texts, despite a certain level of errors, deliver detailed results and insights that are useful for the study of linguistic variation and lexico-grammatical preferences?
  This research question led to the following subquestions, which I mainly address in chapter 1.
  - Can large amounts of parsed data deliver detailed insights into lexical interactions?
    Our initial experiments on corpora with one or 10 million words showed that the data was too sparse, that for the detection of interaction between rare words larger resources were needed. I describe our experiments with using up to 1000 million words in sections 1.2 and 1.3.
  - Can lexical interactions also be found in areas where only few semantic restrictions exist, and in operations that are very productive?
    In areas such as adjective-noun constructions (e.g. *strong tea*), it is well known that strong collocational restrictions apply. It has been less studied whether similar forces are at play in the equally frequent verb-object constructions (e.g. *bridge the gap*) and subject-verb constructions (e.g. *dogs bark*). We

investigate this in section 1.2. Similarly, less productive alternations, such as the dative shift, which we investigate in section 1.3, are subject to many collocational restrictions. In the same section, we apply large corpora to test if the very productive passive alternation, and the Saxon Genitive alternation, also show similar preferences.

  - Is the signal of syntactic variation stronger than the noise produced by parsing errors?
    The counts that one obtains are only reliable if parser errors (the *noise*) do not not overshadow correct output (the *signal*). I investigate in section 1.4.2 if parser errors are forbiddingly influential.
  - Can automatic parsing also be applied profitably to diachronic linguistics, although error rates are higher on texts that are 100 years old or older?
    Error rates are even higher when parsing earlier texts, I therefore investigate improvements in section 1.5.2.
  - Can semantic expectations improve syntactic parsing?
    In addition to lexical preferences like collocations, semantic knowledge can be used to improve parsing, as I explain in section 2.5.5. Until here, I have described how the parser helps in linguistics. From here on, I also show how linguistic insights help the parser. Within limits, the parser can even improve itself, by learning from large amounts of parsed data, in what is called self-training, as I summarise in the same section (2.5.5).

- Can we detect new patterns in language by using data-driven approaches?
  In data-driven approaches, which are sometimes also called corpus-driven, patterns emerge from the data.
  - Can they detect features of variation by genre and region?
    I apply data-driven approaches for the detection of new verb-preposisiton structures in section 1.4.3, using tagged and parsed data, and in section 1.4.4, using chunked data.
  - Can they help us to model the choices that speakers make?
    Just measuring frequencies is often not directly related to choices which speakers make, and this can skew results considerably. I address the question whether data-driven approaches can be used to alleviate this problem in section 1.3.

- Can Descriptive Linguistics profit from insights gained in Text Mining?
  We apply syntactic parsing to Text Mining in chapter 2.
  - Can automatically parsed data, despite a certain level of errors, lead to better Text Mining results?
    This first question led us to the question of how we can detect relevant pattern in the texts. The patterns turned out to be very sparse, as I discuss in section 2.3.
  - Can linguistic insights alleviate the sparse data problem in Text Mining?
    I describe how linguistic insights and data-driven approaches can simplify the patterns in sections 2.3 and 2.4. At the same time, we learn that pure frequencies and discourse features are essential.
  - Can discourse features, such as unity of topic of the document, and the context offered by the document, or the one-sense-per-discourse hypothesis improve Text Mining?
    I dedicate section 2.5 to the question of how important discourse is.
  - Can statistical language models be alternatives to simplistic operationalisations?
    I argue that speaker choices, although they can better approximated with data-driven approaches, are not binary. In order to include more relevant factors, it is essential to move to language models. I argue in section 2.4 that data-driven language models used in Text Mining can give us a more realistic approach to alternations.

- Can language models overcome some of the current shortcomings in Descriptive Linguistics?
  This question is addressed in chapter 3 by investigating the following issues:
  - Can language models answer the shortcomings of significance testing?
    We encounter actual significance testing problems in section 1.4.4, and then discuss additional problems and present solutions in section 3.1. The discourse of a document heavily depends on genres, subgenres and reader expectations, which leads to the following question.
  - Can the shortcomings of local language models be overcome by using a parser as a global model?
    A major reason for using statistical models in linguistics is the multifactorial nature of the speaker choices (e.g. for alternation variants, choice of synonyms, argument structure, collocations), on the one hand in the sense that the interaction between different factors for a single, local choice is enormous. On the other hand, also the choices themselves highly depend on each other, which means that a global system combining distributed interdependent choices may be more appropriate. I argue in section 3.2 that syntactic parsers can take the interactions of the choices at production and disambiguation into consideration. The question which arises from this point is how far parsers can be used as models of human reading:
  - Can automatic parsers potentially be used as psycholinguistic language models?
    In section 3.2 I also show that there is a correlation between psycholinguistic factors and parser-derived measures. I hypothesise that parsers have the potential to be used as psycholingistic language models.

This synopsis aims to summarise the following 16 publications and show the connections that hold between them.

- Descriptive Linguistics (Chapter 1)
  - Collocations (Section 1.2)

    Lehmann and Schneider (2009): Lehmann, Hans Martin and Gerold Schneider. 2009. Parser-based analysis of syntax-lexis interaction. In Andreas H. Jucker, Daniel Schreier, and Marianne Hundt, editors, *Corpora: Pragmatics and discourse: papers from the 29th International conference on English language research on computerized corpora (ICAME 29)*, Language and computers 68. Rodopi, Amsterdam/Atlanta, pages 477–502.

    Lehmann and Schneider (2011): Lehmann, Hans Martin and Gerold Schneider. 2011. A large-scale investigation of verb-attached prepositional phrases. In S. Hoffmann, P. Rayson, and G. Leech, editors, *Studies in Variation, Contacts and Change in English, Volume 6: Methodological and Historical Dimensions of Corpus Linguistics*. Varieng, Helsinki.
  - Alternations (Section 1.3)

    Lehmann and Schneider (2012b): Lehmann, Hans Martin and Gerold Schneider. 2012b. Syntactic variation and lexical preference in the dative-shift alternation. In Joybrato Mukherjee and Magnus Huber, editors, *Studies in Variation, Contacts and Change in English, Papers from the 31st International conference on English language research on computerized corpora (ICAME 31), Giessen, Germany*. Rodopi, Amsterdam.

    Röthlisberger and Schneider (2013): Röthlisberger, Melanie and Gerold Schneider. 2013. Of-genitive versus s-genitive: A corpus-based analysis of possessive constructions in 20thcentury english. In Paul Bennet, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpora*, Korpuslinguistik und Interdisziplinäre Perspektiven auf Sprache - Corpus linguistics and Interdisciplinary perspectives on language (CLIP). Narr Francke Attempto, Stuttgart.
  - Language Variation (Section 1.4)

Schneider and Hundt (2009): Schneider, Gerold and Marianne Hundt. 2009. Using a parser as a heuristic tool for the description of New Englishes. In *Proceedings of Corpus Linguistics 2009*, Liverpool.

Schneider and Hundt (2012): Schneider, Gerold and Marianne Hundt. 2012. "Off with their heads" – profiling TAM in ICE corpora. In Marianne Hundt and Ulrike Gut, editors, *Mapping Univity and Diversity world-wide*, VEAW. Benjamins, Amsterdam.

Schneider (2013): Schneider, Gerold. 2013. Using automatically parsed corpora to discover lexico-grammatical features of english varieties. In Fryni Kakoyianni Doa, editor, *Penser le Lexique-Grammaire, perspectives actuelles. (Papers from the 3rd International Conference on Lexis and Grammar, Nikosia, Cyprus, 5-8 October, 2011).* Éditions Honoré Champion, Paris.

– Language Change (Section 1.5)

Schneider (2012a): Schneider, Gerold. 2012a. Adapting a parser to historical English. In Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen, and Matti Rissanen, editors, *Studies in Variation, Contacts and Change in English, Volume 10: Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*, Helsinki, Finland.

Schneider, Lehmann, and Schneider (2014): Schneider, Gerold, Hans Martin Lehmann, and Peter Schneider. 2014. Parsing Early Modern English corpora. *Literary and Linguistic Computing*, first published online February 6, 2014 doi:10.1093/llc/fqu001.

- Text Mining (Chapter 2)

Schneider, Kaljurand, and Rinaldi (2009): Schneider, Gerold, Kaarel Kaljurand, and Fabio Rinaldi. 2009. Detecting Protein/Protein Interactions using a parser and linguistic resources. In *CICLing 2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 406–417, Mexico City, Mexico. Springer LNC 5449. Best Paper Award, 2nd place.

Schneider, Clematide, and Rinaldi (2011): Schneider, Gerold, Simon Clematide, and Fabio Rinaldi. 2011. Detection of interaction articles and experimental methods in biomedical literature. *BMC Bioinformatics*, special issue on BioCreative III.

Schneider (2012b): Schneider, Gerold. 2012b. Using semantic resources to improve a syntactic dependency parser. In Viktor Pekar Verginica Barbu Mititelu, Octavian Popescu, editor, *SEM-II workshop at LREC 2012*.

Schneider et al. (2012): Schneider, Gerold, Simon Clematide, Gintare Grigonyte, and Fabio Rinaldi. 2012. Using syntax features and document discourse for relation extraction on PharmGKB and CTD. In Fabio Rinaldi, editor, *Proceedings of SMBM 2012*, Zurich.

- Psycholinguistics (Chapter 3)

Schneider and Rinaldi (2011): Schneider, Gerold and Fabio Rinaldi. 2011. A data-driven approach to alternations based on protein-protein interactions. In *Proceedings of the 3rd Congreso Internacional de Lingüística de Corpus (CILC), Valencia, Spain, 7-9 April, 2011.*

Schneider et al. (2005): Schneider, Gerold, Fabio Rinaldi, Kaarel Kaljurand, and Michael Hess. 2005. Closing the gap: Cognitively adequate, fast broad-coverage grammatical role parsing. In *ICEIS Workshop on Natural Language Understanding and Cognitive Science (NLUCS 2005)*, Miami, FL, May 2005.

Schneider and Grigonyte (2013): Schneider, Gerold and Gintare Grigonyte. 2013. Using an automatic parser as a language learner model. In *Book of Abstracts of LCR 2013*, Bergen, Norway, September.

Chapter 1

Applications of Computational Linguistics to Descriptive Linguistics

One of the key instruments for the application of Computational Linguistics to Descriptive Linguistics is my use of syntactically annotated large corpora, which have been annotated automatically. Syntactically annotated corpora have been a desideratum in linguistics for a long time. First, I motivate why syntactically annotated texts are useful: they deliver cleaner results than the application of surface approaches on raw text or part-of-speech tagged texts. If, for example, we wanted to do research on verb subcategorisation and selectional preferences we would need to extract verb-object relations from a corpus. With a surface approach such as regular expressions or an observation window (e.g (Stubbs, 1995)) one would probably incur the following errors.

Precision errors

(1) *Experts fear the virus will spread.*

A surface approach reports verb-object relation between *fear* and *virus*. See Figure1 for a very similar example.

(2) *The report arrived Friday.*

A surface approach reports a verb-object relation between *arrived* and *Friday*.

Recall errors

(3) *John likes swimming.*

A surface approach will not find the verb-object relation, because *swimming* is not a noun, but a verb participle (but conversely including all participles to be objects easily leads to many precision errors).

(4) *John likes, but Mary hates Paul.*

A surface approach will probably not find the verb-object relation, because the distance is quite long.

(5) *The potatoes I like are cold.*

A surface approach will likely not find the implicit verb-object relation in the relative clause. This nonwithstanding, it needs to be pointed out that detecting the object relation is comparatively simple, as it is typically very close to the verb. Subjects, for example, can have relative clauses intervening between the subject and the verb.

When the first corpora with manually added syntactic annotation were published, linguists greeted them with enthusiasm. The first two corpora that have a size of a million words of syntactically annotated text are the Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993) and ICE-GB (Nelson, Wallis, and Aarts, 2002). While these are useful resources, one million words is too few for many applications investigating word-word interactions, because most word types have very few occurrences, as already the famous Zipf's law states (Zipf, 1965). Word-word interaction research, for example on collocations, investigates the co-occurrence of individual words, and thus of rare events in combination. A combination of rare events as in collocations is thus particularly rare, so that often corpora of several hundred million words are needed. It is beyond the scope of any research team to annotate such a resource manually; we need to resort to automatic methods. We will see in section 1.1 that lexical interactions (word-word interactions) have revolutionized our view of grammar.

Automatic parsing has made considerable advances recently and can deliver syntactic annotation for much larger corpora, alleviating sparse data problems and thus opening new perspectives for Descriptive Linguistics. van Noord and Bouma (2009, 37) state that "[k]nowledge-based parsers are now accurate, fast and robust enough to be used to obtain syntactic annotations for very large corpora fully automatically." We apply parsed corpora as a new resource for linguists. Automatically parsed treebanks, also called *parsebanks* or *tree jungles*, have been applied e.g. to Danish (Bick, 2003) and French (Bick, 2010). We use automatically parsed treebanks for several research strands: we use large corpora such as the BNC (Aston and Burnard, 1998) to investigate collocations, as just mentioned, but we also investigate regional English varieties, and historical corpora. No treebanks for English regional varieties or World Englishes exist yet. In this situation, automatically parsed corpora can be used as a stopgap to treebanks. I have parsed the publicly available parts of International Corpus of English (ICE) and many other large corpora like the British National Corpus (BNC) using a dependency parser (Schneider, 2008). The challenges posed by parsing, storing and accessing these large corpora, totalling over 1 billion words, are described in Lehmann and Schneider (2012a).

In my applications to Descriptive Linguistics, I rely on the output that the parser delivers, comparing frequencies reported by the parser. This can only lead to reliable results as long as the parser error rate does not get too high. In order to assess the reliability, detailed evaluations in terms of precision and recall are important. To give a first impression, about 90% of all reported subjects and objects are correct (precision), and over 80% of all subjects and objects in the text are found (recall). I give more evaluations in section 1.4.2, where I will argue that error rates across various genres and varieties are comparable, and spread homogeneously, thus not fundamentally affecting the relative comparison of counts.

1.1 Lexico-Grammar

Descriptive linguistics can profit from syntactically annotated data. Likewise, Computational Linguistics has profited from theoretical linguistics by implementing a wide range of syntactic theories, and from Corpus Linguistics by recognising the central role of statistics. While doing so, it has contributed to revolutionising our view of syntax as an abstract system governed by a closed set of rules.

Sinclair (1991) describes that there are two opposing principles at work in language: the open-choice principle, which is often also called **syntax principle**, in which syntactic rules dominate, and where lexical items can be freely placed in terminal node slots; and the **idiom principle**, in which frequently used word-sequences dominate, and syntactic rules, if they exist at all, are only used rarely and on larger units than local rewrite rules.

> "This [the Open Choice Principle] is a way of seeing language text as the result of a very large number of complex choices. At each point where a unit is completed (a word, phrase, or clause), a large range of choice opens up and the only restraint is grammaticality." (Sinclair, 1991, 109-110).

In rewrite-rule grammars, each rule is local, independent of nodes higher up or further down in the tree, a 'slot and filler' model independent of the lexical material occurring in the terminal nodes. In Sinclair's words, "Any tree structure shows it clearly: the nodes on the tree are choice points." (Sinclair, 1991, 110)

In the idiom principle, "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." (Sinclair, 1991, 110). In a radical version of the idiom principle, there would only be surface sequences of words, and no syntactic abstractions. In a radical version of the syntax principle, all syntactic rules would be equally likely, irrespective of the lexical items, there would either be no syntactic ambiguity, or no way to resolve ambiguity. The idiom principle is the reason why part-of-speech taggers (which are trained on surface sequences) can improve syntactic parsing, and why statistical machine translation typically performs better than syntactic approaches. The syntax principle is the reason why we can construct an infinite number of unseen sentences and makes creative use of syntactic structures, such as multiple embeddings. Obviously the two principles interact:

> syntactic structures and lexical items (or strings of lexical items) are co-selected [...] Particular syntactic structures tend to co-occur with particular lexical items, and – the other side of the coin – lexical items seem to occur in a limited range of structures. The interdependence of syntax and lexis is such that they are ultimately inseparable (Francis, 1993, 143).

It would be useful to assess to which degree each of the two principles are involved. A key aim of scientific endeavour is to find generalisations. Syntactic rules should deliver simple generalisations for language analysis and production. But there are at least two problems complicating this desideratum. **First**, it is often unclear which syntactic rule system should be used. Even simple questions, for example what the top node of a sentence should be, have led to debates. While in early constituency grammars and in Penn Treebank annotation, S is the top node, in Dependency Grammar (Tesnière, 1959) the top node is a verbal projection, and in many versions of Principles and Parameters, it is a projections of the typically empty CP. In Dependency Grammar (DG), locality extends to the clause level (Schneider, 2005) like in Tree-Adjoining Grammar (Frank, 2002), which has the potentially positive effect that fronting does not posit a movement (or secondary link in DG terms, or co-indexation in HPSG terms), and the potentially negative effect that in DG there is no distinction between inner arguments (such as objects, which only transitive verbs have) and outer arguments (the subject). From a practical viewpoint, the differences between the various models are often unconsequential.

Other insights on the nature of grammar, largely aided by the advent of Corpus Linguistics and computational tools, have affected all syntactic theories alike in a more fundamental way: the realisation that syntactic rules are largely lexicon-driven, and that lexical interactions are pervasive. While Sinclair (1991, 110) writes that "[v]irtually all grammars are constructed on the open-choice principle" and that "[t]his is probably the normal way of seeing and describing language" (Sinclair, 1991, 109) most theories have since adopted a radically more lexico-centric approach. This is the **second, and more consequential** complication of our desideratum to be able to rely on a simple and closed set of syntactic rules. The insight that grammar is largely lexicon-driven comes from a number of sources: syntactic parsing in Computational Linguistics, grammar theory, Corpus Linguistics and Psycholinguistics.

In syntactic parsing, the blind application of syntax rules often leads to hundreds of analyses for real-world sentences. For disambiguation, bi-lexical preferences (Collins, 1999) can be used. For example, the parser Pro3Gres (Schneider, 2008) uses maximum-likelihood estimation (MLE) to estimate the probability the the dependency relation R is found at distance (in chunks) $dist$, given the lexical head a of the governor and the lexical head b of the dependent.

$$p(R, dist|a, b) = P(R|a, b) \cdot P(dist|R, a, b) \cong \frac{f(R, a, b)}{f((\sum R), a, b)} \cdot \frac{f(R, dist)}{fR} \tag{1.1}$$

The object dependency rule, for example, states that a noun preceded by a verb is licensed to be attached as object. The adjunct dependency rule is very similar, and often in conflict with the object rule. If the verb is *eat* and the noun is *pizza*, then the probability $p(obj|eat, pizza)$ is high, while the adjunct probability is very low. If the verb is *eat* and the noun is *yesterday*, the probability $p(adjunct|eat, yesterday)$ is high, while the object probability is very low.

In formal grammar, explicitly lexicalist theories like Lexical-Functional Grammar (LFG) and Head-Driven Phrase Structure Grammar (HPSG) gained importance, and in Chomskyan approaches syntactic structures are projected from the lexicon, constrained solely by syntactic skeletons like X-bar structures. It is further recognised, for example in Kaplan et al. (2004), that deep-linguistic formal grammar parsers, without using results from word sequence statistics gained from part-of-speech tagging, and word interaction statistics gained from probabilistic parsing, cannot achieve the robustness, coverage, accuracy and speed required for large-scale application.

In Corpus Linguistics, studies in Sinclair's tradition such as Altenberg (1998) and Moon (1998) have investigated whether, in addition to the well-known strong lexical preferences in classical idioms, language is full of partly prefabricated structures, phraseological expressions, and frequent word combinations. "These frequently occurring multi-word expressions, which tend to go virtually unnoticed in everyday language because they are not very salient psychologically, seem to be used as the usual or preferred building blocks in speech and writing." (De Cock, 2000, 51). Lehmann and Schneider (2009) and Lehmann and Schneider (2011), which I discuss in section 1.2, can also be placed in this tradition. Hoey (2005, 1) goes as far as to claim that "lexis is complexly and systematically structured and that grammar is an outcome of this lexical structure", which would entail a completely lexis-driven approach, in which syntax is just an epiphenomenon.

In psycholinguistics, it has been noted that our creativity in the use of syntactic constructions is very restricted. In language acquisition, it has been shown that the vast majority of chidrens' early language uses item-based linguistic schemas (Tomasello, 2000). Children exclusively use the lexical-specific syntactic formats with which they have started the acquisition of their native language prior to developing abstract schemas or rules. The idiom principle is learnt first, the syntax principle later. Ninio (2006) argues that as a consequence for grammar research, syntax is lexicon-driven. We can thus expect to see strong lexical interaction across all syntactic rules. There thus are indications that we may answer positively to the research question *Can lexical interactions also be found in areas where only few semantic restrictions exist, and in operations that are very productive?* Pawley and Syder (1983) argue that the crucial difference between native English and non-native English is that non-native speakers do not master the subtle phraseological preferences such as collocations to the same degree. As lexico-grammatical differences are subtle, leading only to small statistical trends, and as most words are rare, we need large amounts of annotated data. The following studies investigate these issues further.

- detection of collocations (Lehmann and Schneider, 2009; Lehmann and Schneider, 2011), see section 1.2
- detection of the envelope of variation in alternations (Lehmann and Schneider, 2012b; Röthlisberger and Schneider, 2013) see section 1.3
- verb-preposition constructions variation (Schneider and Zipp, 2013; Schneider, 2013), see section 1.4
- regional differences in the morphosyntactic features tense, aspect and modality (TAM) (Schneider and Hundt, 2012), see section 1.4
- diachronic linguistics (Schneider, 2012a; Röthlisberger and Schneider, 2013; Schneider, Lehmann, and Schneider, 2014), see section 1.5
- zero determiners variation (Schneider, 2013), see section 1.4
- detection of transparent words (Schneider, Kaljurand, and Rinaldi, 2009), see section 1.6
- alternations as an open set (Schneider and Rinaldi, 2011), see section 1.6

In my work, I follow the general tenets of Construction Grammar, e.g. Stefanowitsch and Gries (2003). I use large amounts of parsed data, between 300 and 1000 million words. This means that we can reduce sparse data problems in some areas, particularly if lexis is involved, as most lexeme are rare (Zipf, 1965). Because of the reduction of sparseness, large-scale parsing helps realistic Construction Grammar, as the data is more reliable and less affected by random fluctuations. In these approaches, one measures the signal that the parser emits, helped by some data-driven methods.

The insistence on lexical interactions has ugly consequences. It leads to enormous amounts of idiosyncratic data. When John Sinclair writes "trust the text" (Sinclair and Carter, 2004) he also means trust the data in its complexity and resistance to simplification. Data compression is only possible as long as data loss does not add an important skew. In this synopsis, and much more so in the papers connected to it, I often give long lists, and many graphs, illustrating the complexities of the data, which I do not want to hide from the reader. The complexity of the data, which stems form a large, potentially infinite number of sources (or what one calls *factors* or *features* in statistics) is a main motivating factor for the use of models in chapter 2: they can pay respect to the many interconnected factors that come into play. If not only the factors are interconnected, but also the outcomes to which these factors lead, then one needs more complex models. This is what I discuss in chapter 3.

1.2 Collocations

A major emanation of strong word-word interactions as we find them according to the idiom principle are collocations. A collocation is defined, according to Choueka (1988, 609), as "a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components." Criteria for the definition of collocation typically include the following.

- Non-compositionality (e.g. *kick the bucket*)
- Non-substitutability. Near-synonyms cannot be used (e.g. **yellow wine*)
- Non-modifiability: *kick the bucket*, but **kick the buckets, *kick my bucket*
- Non-literal translations: *red wine* ↔ *vino tinto*, *take decisions* ↔ *Entscheidungen treffen*
- statistical co-occurrence

The last criterion, co-occurrence, can be investigated quite easily with frequency-based computational approaches, with association measures from significance testing and Information Theory. Non-modifiability can also be investigated quantitatively. Non-substitutability is more difficult to measure. But according to Fernando and Flavell (1981, 17), semantic non-compositionality is the touchstone for idiomaticity.

Wulff (2008) compares two methods: substitution-based approaches, which test for substitutability, and a similarity-based approach, which uses association measures.

- Substitution-based approaches measure compositionality by means of substitutability: given a potential collocation containing the word W, how likely is it that the word W' which is a synonym of W, can appear instead of W? Large synonym dictionaries can be used to generate and test W'. Wulff (2008) uses three recent substitution-based methods (Lin, 1999; Schone and Jurafsky, 2001; McCarthy and Carroll, 2003).
- Similarity-based approaches measure associations between the words in a potential collocation.

Although the substitution-based method intuitively seems like a better proxy to the the touchstone of non-compositionality than the similarity-based method, Wulff (2008) finds the similarity-based approach performs better. This indicates that frequency-based approaches can be a good stopgap to inherently semantic tasks.

I often use O/E as collocation measure. O/E, literally defined as Observed divided by Expected, is a measure of surprise. It delivers the same ranking as mutual information (MI). The Expected probability is the chance occurrence of the two words in the collocation pair (x,y) given their individual probabilities, i.e. it expresses the probability that the events happen independently. The Observed probability is the joint probability of x and y appearing together (in the same constriction, or inside an observation window) as seen in the data. O/E is defined as:

$$O/E = \frac{P(x,y)}{P(x) \cdot P(y)} = \frac{f(x,y) \cdot N \cdot N}{N \cdot f(x) \cdot f(y)} = \frac{f(x,y) \cdot N}{f(x) \cdot f(y)} \tag{1.2}$$

where N is the size of the corpus. O/E has good recall on rare collocations, but has a tendency to over-report rare collocations (Evert, 2009): in traditional windows-based approaches, false positives dominate the highest ranked items. But approaches based on parsed corpora provide considerably cleaner data (Seretan and Wehrli, 2006; Seretan, 2011). O/E delivers very good results on parsed data, when it is paired with a T-score significance threshold. T-score is defined as (see Evert (2009))

$$T = \frac{O - E}{\sqrt{O}} \tag{1.3}$$

In some of my investigations, I have also included a non-modifiability measure. I use Yule's K (Yule, 1944) and (Malvern et al., 2004, 44), which is defined as

$$K = 10^4 \cdot \frac{(\sum_{r=1}^{N} v_r r^2) - N}{N^2} \tag{1.4}$$

where v_r is the number of types which occur r times in a text of length N. For example r is equal to 1 for hapax legomena. Low K indicates high variability.

Since most word types have very few occurrences (Zipf, 1965) lexis-lexis interaction data is typically very sparse. An investigation of collocation requires large amounts of data, however. For this reason, I have used corpora containing between several hundred million and thousand million words.

1.2.1 Subject and Object Collocations

While subject-verb and verb-object collocations may appear to be less prototypical candidates for collocations than adjective-noun pairs (e.g. *strong tea*, *powerful car*) or verb-PP constructions (e.g. *bear in mind*), there are at least some verb-object collocations like the proverbial *kick the bucket*. Also, they are marked by a high degree of selectional restrictions, word-word combinations are frequent due to semantic reasons.

subject_verbhead	f(subj_verb)	f(subject)	f(verb)	O/E
tentacle_pore	77	136	243	8.78e+10
onion_chop	56	139	776	1.95e+10
egg_hatch	58	396	456	1.21e+10
doorbell_ring	65	80	4220	7.26e+09
interview(s)_record	136	136	5389	6.99e+09
bomb_explode	158	652	1326	6.89e+09
rumour_circulate	56	402	848	6.19e+09
telephone_ring	195	356	4220	4.89e+09
dog_bark	81	1739	506	3.47e+09
phone_ring	142	374	4220	3.39e+09
relation_deteriorate	62	950	738	3.33e+09
sun_shine	294	1981	1690	3.31e+09
lip_part	63	825	872	3.30e+09
lifespan_display	111	420	3391	2.93e+09
god_bless	135	3349	603	2.52e+09
wind_blow	239	1381	2986	2.18e+09
thief_steal	103	590	3015	2.18e+09

Table 1.1: Subject-Verb collocations, sorted by O/E

E	verb_object	f(verb_obj)	f(verb)	f(object)	O/E
	inter_alia	259	348	269	8.26e+10
+	wreak_havoc	88	157	248	6.74e+10
+	whet_appetite	70	85	419	5.86e+10
	rick_sky	83	91	500	5.44e+10
+	extol_virtue	56	132	379	3.34e+10
	programme_tdy	55	1068	55	2.79e+10
+	clench_fist	82	399	403	1.52e+10
+	beg_pardon	145	1320	216	1.51e+10
+	grit_tooth	146	227	1363	1.40e+10
+	purse_lip	135	184	1680	1.30e+10
+	wrinkle_nose	82	202	1039	1.16e+10
+	bridge_gap	162	321	1367	1.10e+10
+	sow_seed	107	469	637	1.06e+10
+	heave_sigh	74	512	430	1.00e+10
	buck_trend	58	184	1004	9.37e+09
	enclose_sae	68	1148	211	8.38e+09
	ratify_treaty	99	419	859	8.21e+09
+	reap_reward	73	394	680	8.13e+09

Table 1.2: Verb-Object collocations, sorted by O/E

The strongest subject-verb collocations according to O/E, which are reported in Lehmann and Schneider (2009), are given in Table 1.1. They are dominated by semantic facts such as *wind blow* and *egg hatch*, but also contain tagger errors such as *tentacle pore*. The strongest verb-object collocations found in Lehmann and Schneider (2009) are given in Table 1.2. They contain semantic facts such as *sow seed*, tagger errors like *inter alia* and strong collocations like *wreak havoc*, *whet appetite*, *heave sigh*, *reap reward*, where the same semantic concepts could equally be (but typically aren't) expressed by different verb-noun combinations. As Hoey (2005) predicted, also subject and object relations are subject to strong collocations, lexical priming occurs everywhere. The large amounts of data that our investigation delivered can also be used in lexicology to establish the syntactic frame of a verb, and its selectional restrictions. In Computational Linguistics, the data can be used for the construction of distributional semantic models (Curran, 2004; Rothenhäusler and Schütze, 2009; Grefenstette et al., 2011), where items sharing the same syntactic contexts are considered near-synonyms. I will discuss some of these approaches in chapter 3. Key insights in Lehmann and Schneider (2009) also include an evaluation on BNC data, which shows that the performance of my parser (Schneider, 2008) is not considerably lower on BNC written texts than on the training domain of newspaper texts.

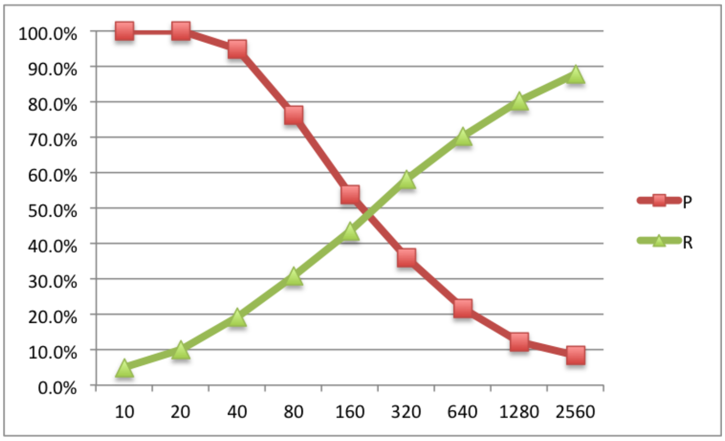


Figure 1.1: Precision (P) and recall (R) of light verb constructions with the verb *give*, according to the T-Test collocation measure, relative to the position in the candidate list

Going beyond Lehmann and Schneider (2009) I add a little evaluation. The first column in Table 1.2 (E, for Evaluation) again contains my assessment on whether the collocation candidate is a collocation: if I agree, it is marked as '+'. Precision for the top of the list is 12/18, about 66%. Precision decreases further down in the list. The assessment of recall is more involved, as it would require a gold standard to which one can then compare. Ronan and Schneider (submitted) have investigated light verb constructions. Light verb constructions are verb-object combinations in which the semantics of the construction is non-compositional or depends more on the object than on the verb. An example is *heave sigh*, which also has a simplex verb correspondance *sigh*. We have constructed a small gold standard. A complete gold standard would require to know all collocations, which is basically impossible, as they form an open list. In the case of light verb constructions, the possibility of being able to compile one is increased if the light verbs are constrained to a closed list (which is a possible oversimplification for the sake of operationalisation). In our approach, we have used a manually compiled, long list (200 entries) of one prototypical support verb, *give*, to approximate an estimate for recall. Figure 1.1 plots precision and recall, relative to the position in the list of light verb candidates, on a logarithmic scale. Precision starts off at 100% for the first 20 positions in the list, and drops to below 10% around position 2560, reaching a recall of almost 90%. We have applied the T-score collocation measure and a nominalisation lexicon to the parsed output of the 100 million word British National Corpus (BNC). The fact that even at the end of the list precision and recall have not

E	verb	prep	desc noun	modification K	derminers K	t-score	O/E	modifiers
+	pale	into	insignificance	8787.5	9750	6.32	387428	bland relative
	contain	within	begins	9722.22	9722.22	5.99	310203	box
	infect	with	hiv	9807.69	9430.47	7.21	64602.1	-
+	breathe	down	neck	9729.73	9729.73	6.08	43999.3	-
+	mutter	under	breath	9743.59	9743.59	6.24	33961.9	-
+	burst	into	tear	9721.37	9906.54	10.34	18031.1	noisy
	summarise	in	a	9918.03	9918.03	11.04	13981.4	appendix
+	roar	with	laughter	9843.75	9843.75	7.99	11577.2	-
+	hope	against	hope	9714.29	9159.18	5.91	11546.6	-
+	sigh	with	relief	9262.5	9750	6.32	9674.92	silent
+	gasp	for	breath	9836.07	9836.07	7.80	6590.54	-
	be	if	anything	9736.84	9736.84	6.16	5456.4	-
+	obtain	by	pretence	9615.38	9615.38	5.09	5346.81	false
	sue	for	damage	9743.59	9743.59	6.24	5125.58	-
	be	en	route	9761.9	9761.9	6.47	5099.94	-
	feel	like	cry	9629.63	9629.63	5.19	5001.65	-
	give	up	smoking	9391.86	9876.54	8.99	4879.38	cigarette drinking
	screw	up	eye	9313.14	9767.44	6.55	4677.19	cornflower
+	fall	into	disrepair	8954.08	9642.86	5.29	4615.43	disuse
	mention	in	subsection	9161.71	9161.71	7.61	4297.24	subsection
	glance	at	watch	9330.82	9565.01	15.77	4262.36	gold spiderman small fob ancient
	pick	up	receiver	9183.33	8883.33	7.74	3659.22	dangling telephone
+	start	from	scratch	9876.54	9876.54	8.99	3163.1	-

Table 1.3: VPN tuples from BNC, ordered by O/E, *filtered for low variability* on determiners and modifiers with Yule's K

flattened out indicates that even for this restricted task one needs to expect an open list, i.e. an infinite number of light verb constructions with the verb *give*. When including all verbs, precision at position 50,000 is still around 5%, again suggesting an open list. I have also evaluated using O/E or T-score on the ICE-corpora instead of BNC. Results were considerably worse, which confirms that very large amounts of data are required for the data-driven detection of collocations.

1.2.2 Verb-PP Collocations

Verb-PP combinations form prototypical collocations. We have investigated them in Lehmann and Schneider (2011), looking at verb-preposition-noun (VPN) and verb-object-preposition-noun (VOPN) combinations. The written part of the BNC, which contains 90 million words, returns 5,156,281 VPN triple tokens resulting in 1,129,643 VPN triple types in our dependency parsing approach. Compared to all possible VPN combinations in verb-attached PP structures, only one in a million possible VPN combinations ever occurs. The distribution is highly Zipfian: the most frequent triple occurs over 2,000 times, while 896,963 types (about 80%) are hapax legomena, i.e. they occur only once. These findings confirm Pawley and Syder (1983, 193)'s observation "that native speakers do not exercise the creative potential of syntactic rules to anything like their full extent".

In addition to measuring the collocation strength with O/E we also measure modifiabillty with Yule's K (Yule, 1944), on the choice of determiners and pre-modifying adjectives and nouns. Idioms and collocations are typically more fossilized and show less variation. A sample result from the many more results in the paper is given in Table 1.3. Key insights in Lehmann and Schneider (2011) also include the recognition that a realistic approach to construction grammar (Stefanowitsch and Gries, 2003) requires large amounts of data. The first column in Table 1.3 (E, for Evaluation) again contains my assessment on whether the collocation candidate is a collocation, marked by a '+'. Precision for the top of the list is 11/23, about 50%. If one leaves away the filter for high Yule's K, precision is considerably lower. I give only the top 23 lines here to serve as examples. Lehmann and Schneider (2011) is an online publication, which allowed us to furnish full lists, often hundreds to thousands of lines, allowing the reader to browse the whole gradience from idiomatic to lexical preference to free selection and explore a data-oriented linguistics approach.

subject_verbhead	f(subj_verb)	f(subject)	f(verb)	O/E
seed_sow	51	172	147	1.67e+09
shot_fire	89	233	474	6.69e+08
lesson_learn	73	217	604	4.62e+08
duty_owe	59	431	282	4.03e+08
battle_fight	55	235	521	3.73e+08
breakfast_serve	55	113	1375	2.94e+08
offence_commit	193	398	1459	2.76e+08
warrant_issue	73	150	1511	2.67e+08
day_number	62	989	207	2.51e+08
power_vest	124	1564	274	2.40e+08
battle_win	53	235	788	2.37e+08

Table 1.4: Top-ranked Subject Passive Verb Collocations

1.3 Alternations

Alternations are defined as syntactic variations exploiting the fact that "language can provide different syntactic means to express the same propositional content, to convey the same information" in Kreyer (2010, chapter 9: p. 169). Alternations have been extensively researched. Levin (1993) distinguishes 79 different alternations. The best investigated alternations include the passive voice alternation (e.g. (Leech et al., 2009, chapter 7) and (Seoane, 2009)), the dative shift (e.g. (Mukherjee, 2005; Bresnan and Nikitina, 2009)) and the Saxon Genitive alternation (e.g. (Jucker, 1993; Rosenbach, 2002; Kreyer, 2003)). I summarise our research on these alternations, and then address some criticisms in section 1.3.4.

1.3.1 Passive Alternation

The passive alternation is particularly productive and, though intuitively the alternation may seem largely unconstrained, clear patterns can be determined. We have investigated verb-passive subject collocations and alternations in Lehmann and Schneider (2009) using several hundred million words of running text. In Table 1.4 I give the top ranked collocations ranked by O/E. The list contains idioms like *day number* as in *his days are numbered.* Table 1.5 compares passive constructions in comparison to their active counterparts. Such an investigation requires the identification of thousands of instances of syntactic active and passive subjects with good precision and recall, it would thus be very difficult to do such investigations without an approach based on parsing. While the passive construction is generally an order of magnitude less frequent than the active construction and depends on the genre (see section 1.4), Table 1.5 shows that certain verbs and particularly certain verb-subject combinations have a strong preference for the passive.

The top-ranked item for passive preference, *baby_bear* as in *babies are born* is an interesting case, as the active and the passive split almost coincides with the two main verb senses: *give birth* (which is almost always in the passive voice) and *bear something*, which is almost always in the active, and is dominated by the construction *bear something in mind.* In most other cases, there is a pragmatic, but no fundamental semantic difference between the passive and its corresponding active construction. In order to determine the envelope of variation (Labov, 1969; Sankoff, 1988), comparing all active and passive constructions is thus a reasonable approximation. Let us elaborate on the concept of envelope of variation in the following sections.

1.3.2 The Dative Alternation and its Envelope

I have just argued that for the passive alternation, comparing all active and passive constructions is a reasonable approximation. For the dative alternation, which is often called *dative shift*, the situation is totally different. Counting all verb-double object (such as *give him a book*) and verb-*to*-PP constructions (such as *give a book to him*) would include many instances that are not in the envelope of variation (Labov, 1969; Sankoff, 1988), for

pair of lemmas	f(active)	f(passive)	f(total)	% passive
baby_bear	3	141	144	97.91
study_carry	4	118	122	96.72
committee_set	5	137	142	96.47
power_vest	6	124	130	95.38
test_carry	7	100	107	93.45
research_carry	10	125	135	92.59
system_base	10	106	116	91.37
work_carry	29	274	303	90.42
example_show	47	253	300	84.33
case_adjourn	23	94	117	80.34
election_hold	112	442	554	79.78
people_arrest	31	113	144	78.47

Table 1.5: Subject Passive Verb Preferences

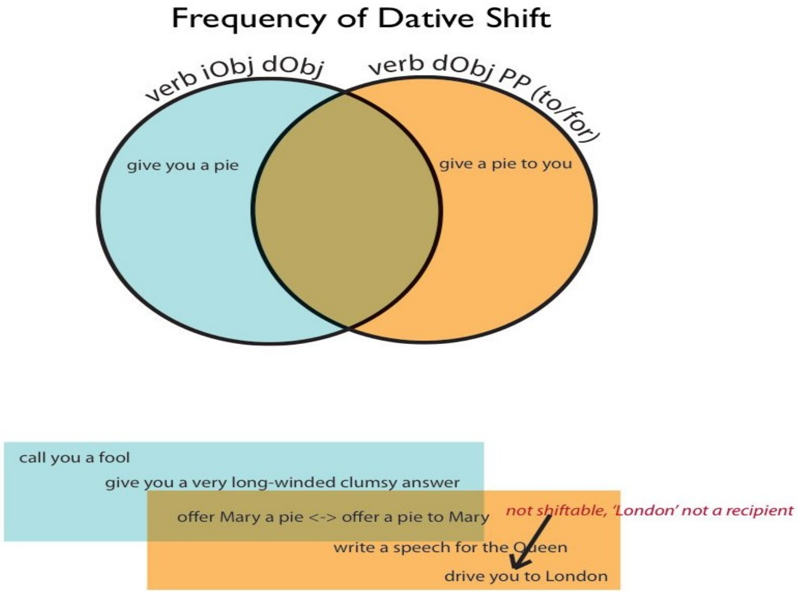

Frequency of Dative Shift
verb iObj dObj
verb dObj PP (to/for)
give you a pie
give a pie to you
call you a fool
give you a very long-winded clumsy answer
offer Mary a pie <-> offer a pie to Mary
not shiftable, 'London' not a recipient
write a speech for the Queen
drive you to London

Figure 1.2: Constraining the envelope of variation by means of lexical overlap

example *drive him to London*. The "envelope of variation" describes if a construction can be alternated or not: *give a book to him* can be alternated to *give him a book*, but *drive him to London* cannot be alternated to **drive London him*. Further restrictions are thus necessary. I discuss in the following how one can learn a crucial restriction using a corpus-driven method, i.e. learn directly from the data. The method is described in Lehmann and Schneider (2012b).

The distinction between corpus-based and corpus-driven methods was introduced in Tognini-Bonelli (2001). In corpus-based approaches, existing hypothesis are tested, in corpus-driven or data-driven approaches, hypotheses arise from the corpus data.

Corpus-driven approaches have advantages and disadvantages. An advantage is that, in areas of gradience and subtle differences, they can bring patterns to the surface that went unnoticed by linguists (e.g. Hunston & Francis, 2000), as variationist linguistics is often subtle and gradient. A disadvantage and potential danger of corpus-driven approaches is that they depend directly on the corpus and its sampling: "... since the information provided by the corpus is placed centrally and accounted for exhaustively, then there is a risk of error if the corpus turns out to be unrepresentative" (Tognini-Bonelli, 2001, 88). For corpus-driven approaches, large amounts of data are necessary, and relying on frequencies implies a tacit hypothesis, namely that significant frequency differences in

lemma triplet	dshift	to	for	% dshift	iObj
ask you question	4876	3	8	99.8	you
tell you truth	1203	4	1	99.6	you
tell you story	958	3	3	99.4	you
ask him question	1089	6	1	99.4	him
show you picture	1698	13	1	99.2	you
give you number	470	3	1	99.2	you
bring you update	456	5	0	98.9	you
give them information	519	6	0	98.9	them
bring them home	502	6	0	98.8	them
ask them question	404	3	2	98.8	them
...		...			...
cost you price	4	0	4	50	you
send us e-mail	19	19	0	50	us
give administration mark	4	4	0	50	administration
give bush vote	5	5	0	50	bush
send them question	8	8	0	50	them
do himself damage	6	5	1	50	himself
send russia message	5	5	0	50	russia
show audience picture	4	4	0	50	audience
owe him life	7	7	0	50	him
issue us statement	9	9	0	50	us
give child drug	5	5	0	50	child
...		...			...
send country message	4	42	0	8.7	country
give company break	9	99	0	8.3	company
do you something	11	35	90	8.1	you
send people message	20	279	3	6.6	people
send child message	4	62	0	6.1	child
do you anything	10	49	143	5	you
send world message	5	148	0	3.3	world
do you work	4	2	121	3.1	you
pay it attention	14	587	1	2.3	it

Table 1.6: (a) Dative shift preferences.

verb	sum(f_pp)	sum(f_dshift)	% PP
make	74723	67	99.91
get	154971	258	99.83
extend	2074	191	91.56
charge	3638	415	89.76
appoint	451	70	86.56
deliver	3107	542	85.14
vote	3084	538	85.14
predict	205	45	82
lend	1059	285	78.79
assure	144	40	78.26
mail	238	68	77.77
choose	734	213	77.50
award	411	121	77.25
serve	7539	2696	73.65
elect	572	255	69.16
sell	7616	3501	68.50
pay	22805	12539	64.52
provide	14156	7795	64.481
count	675	372	64.46
...			
bring	48508	107643	31.06
win	1458	3515	29.31
grant	13	34	27.65
ask	18728	52301	26.36
tell	39965	129902	23.52
answer	473	1683	21.93
cost	275	1472	15.74
guarantee	74	449	14.14
call	16016	131795	10.83
show	2917	33761	7.95
give	32645	432847	7.01

(b) Dative shift preferences summed per verb the investigated data are indicative.

The above quote (Tognini-Bonelli, 2001, 88) also entails that inaccurate operationalisations, such as an overly simplistic envelope of variation can lead to skewed results. I suggest to use a data-driven way to constrain the envelope of variation: if instances for both constructions can be found with the same lexical participants, one accepts both as being each other's alternate variant, otherwise one discards them. Our lexical overlap method is illustrated in Figure 1.2: the complements of copular verbs and *elect* verbs are conceptualised by the parser as objects. As the double object construction *call you (a) fool*, which occurs in the corpus, does not find a verb-of PP counterpart *call (a) fool to you* (the lexical participant triple *(call,you,fool)* is absent for the to-PP variant) it is excluded from the envelope of variation. The second example, *give you a very long-winded clumsy answer* is possible, but it is relatively unlikely to find its counterpart *give a very long-winded clumsy answer to you* as the factors pronominality and end-weight favour the double object construction. Due to sparseness, we only measure lexical overlap of the head lemmas, and we did not include length as a constraining factor as this would include circularity into the procedure. The danger of running into a circular argument is low for our lexical overlap method, as it is a semantic method which only takes the assumption that two words typically mean the same thing, even if they occur in different contexts, grammatical functions, and morphological forms. Accordingly *give you a very long-winded clumsy answer* finds a counterpart *give answer to you* (lexical participant triple *(give,you,answer)* is found in both configurations). The third example, *offer Mary a pie* will likely find a verb-to PP counterpart. *Write (a) speech for (the) Queen* is a benefactive construction, which can also be in the envelope of variation. The concrete example here is intuitively unlikely to find a counterpart *write (the) Queen (a) speech*, the data decides for us – indeed there is no counterpart in our data. The last example, *drive you to London*, also finds no counterpart *drive London you* and is thus rightly excluded from the envelope.

The preferences in the dative alternation can be seen in Table 1.6 (a) per lexical construction, and summed over each verb in Table 1.6 (b). For space reasons, only excerpts are given. We see, for example, that the top of Table 1.3.2 is dominated by constructions which have a pronoun as indirect object, showing that pronominality is a strong factor. The middle of the lists, where the probability for both constructions is around 50%, is quite small. If the choice were random, most constructions would cluster in the middle an form a normal distribution. This is obviously not the case: factors like pronominality, animacy, and lexis play crucial and almost deterministic roles, as Bresnan et al. (2007) have shown. The detailed gradient descriptions that these tables offer can be used for lexicography, or for a regression model similar to Bresnan et al. (2007), with the difference that we can offer a much larger dataset, while they used very rich semantic factors in addition to lexis.

1.3.3 Saxon Genitive Alternation

I have used a similar approach for describing the Saxon Genitive Alternation in Röthlisberger and Schneider (2013). In particular, I have also used lexical overlap to constrain the envelope of variation. As I have also used the approach for a diachronic description, I will come back to it in section 1.5.

1.3.4 Criticism and Outlook

In a classical approach, alternations are two syntactic configurations that are used to convey the same meaning. The following 8 illustrative examples are adapted from Schneider and Rinaldi (2011, 4), which I follow closely here.

The application of alternations is subject to many restrictions. They are known as the envelope of variation (Labov, 1969; Sankoff, 1988) or the choice context (Rosenbach, 2002). They rule out contexts in which speakers do not have a real choice between the two variants. For example,

(6) Peter gave a book to those students who had achieved a grade A mark.

is acceptable, but 7 will hardly ever be used:

(7) ?Peter gave those students who had achieved a grade A mark a book.

Sentence 7 is highly marked as it violates the the principle of end weight, the linguistic tendency to put short constituents first, and long constituents later. Similarly, while

(8) Mary's picture of the house is great.

is acceptable, 9 is highly marked,

(9) ?Mary's house's picture is great.

because nested Saxon genitive constructions are typically avoided, and also because the *of*-PP in 8 does not necessarily express a possessor relation, which is the prototypical function of the Saxon Genitive.

For each alternation, there are over a dozen such restrictions, including idioms (e.g. *point of view, Noah's ark, earth's crust*), descriptive genitives (e.g. *woman's magazine*), measures (e.g. *a tin of soup*), quality genitives (e.g. *image of power*), see e.g. Jucker (1993).

Often, only a minority of all candidate configuration tokens are really available for the alternation. We have manually assessed for the Saxon Genitive alternation whether the alternation is available. On 307 Saxon Genitives, 259 (84%) could also have been rendered as an of-PP without a strong change in meaning or rendering the sentence unacceptable. On 305 of-PPs, only 116 (38%) could equally have been expressed with a Saxon Genitive. It is often difficult to explain why one variant is not used, e.g why is *one's health* used but not *health of one*, why is *concentration of oxygen* used but not *oxygen's concentration* ?

While syntactic restrictions can be listed, the set of semantic restrictions is possibly infinite. The verbal semantics of *give*, for example entail that sentences 10 and 11 are only equivalent if the printout of a speech is the topic of conversation.

(10) Mary gave a speech to the students.

(11) Mary gave the students a speech.

The deep-syntactic role typically depends on verb semantics. In the nominalization alternation, for example, *destruction of the city* implies *city* as object, while in *implication of the discovery* the word *discovery* is a subject. Such behaviour can be found in most alternations. For example, *God's creation* and *the creation of God* are probably not in the envelope of variation.

We have shown in (Lehmann and Schneider, 2012b) that non-core ditransitive verbs have a different behaviour from prototypes. The most prototypical verb, *give*, has a preference for the double-NP construction, while marginal ditransitives such as *provide*, are rarely used with the double-NP construction. It is unclear if a list of ditransitive verbs can be compiled in the first place. There are indications that they form an open class.

Some verbs, for example *provide*, illustrate how the alternation can take many forms. The double-NP construction is not the alternative form to an NP + *of*-PP construction, but to an NP + *with*-PP construction, or an NP + *to*-PP construction. The double-NP construction (e.g. 12) is rare, the BNC only contains a few dozen of double-NP constructions, about 4000 with-PP constructions (e.g. 13), and about 2000 to-PPs (e.g. 14).

(12) You provided him his death, others have provided him a grave. (BNC-Wri K8S)

(13) The forwards played extremely well as a unit, driving in unison and providing their backs with good ball. (BNC-Wri K5A)

(14) Salespeople may also be called upon to provide after-sales service to customers. (BNC-Wri K94)

When including the benefactive construction, (e.g. *bake Mary a cake* vs. *bake a cake for Mary*) the alleged binariness generally collapses. In those cases where both for-PP and to-PP are available, the meaning is often completely different (*do something for someone* vs. *do something to someone*)

In order to address these aspects of alternations, instead of viewing alternations as solely a binary decision between two choices a view of alternations as a multifactorial phenomenon of many choices, relating the many different ways of expressing similar concepts to each other would be more appropriate. I have used a data-driven approach to constrain the envelope based on lexical overlap of the constructions, but a data-driven approach to the non-binariness of alternation is still lacking. While viewing alternations as binary is a suitable operationalisation, it is fundamentally inappropriate, as e.g. Arppe et al. (2010) express.

> "Our focus on alternations is the result of theoretical heritage from generative syntax and a matter of methodological convenience. Most linguistic decisions that speakers make are more complex than binary choices ... alternations are as simplistic and reductionistic as the theories of language that originally studied them" (Arppe et al., 2010).

Let us take up these points again in section 2.4.

1.4 Language Variation

Language variation by region and by genre is a major area of corpus research. I first give an overview of measuring obvious genre differences reported by the output of the parser (section 1.4.1), showing that genre variation is marked. I then discuss how far one can trust the parser signal (section 1.4.2). As regional variation is often too subtle to deliver signal differences, I give a case study of a data-driven approach to regional variation (section 1.4.3). Data-driven approaches force the researcher to interpret many features, which possibly interact with each other. I give an example (section 1.4.4) and show typical shortcomings of significance testing. I conclude that my findings provide further evidence for the suggestions by Evert (2006), Gries (2010) and Gries (2012) that predictive language models deliver more accurate descriptions.

1.4.1 Genre Variation

As text genre is one of the most critical parameters influencing linguistic variation, let us start by giving a few examples of variation by genre. I use 10,000 random sentences from BNC for each genre, and I apply the genre classification from David Lee's detailed BNC classification (Lee, 2001). From the vast array of syntactic differences, I illustrate genre differences in NP complexity, namely postmodification by PPs, and frequency of relative clauses.

Figure 1.3 shows PP-attachment frequencies per verb-centre. We see that PPs attached to nouns are more than three times as frequent in the academic and scientific genre than in fiction and prose. The difference is smaller for PPs attached to verbs. Figure 1.4 shows frequencies of postmodification by relative clause, measured per verb-centre. The dependency label $modrel$ attaches a full relative clause to a noun, the label $modpart$ attaches reduced relative classes (participles) to a noun. While full relative clauses are as frequent in fiction and prose as in scientific texts, reduced relative clauses are extremely frequent in the scientific genre, probably as they are an efficient way to compress information (Leech et al., 2009). These findings underline that text genre must be accorded a central status in each discussion of language variation. We have also investigated the historical development of relative clauses in a diachronic perspective in Hundt, Denison, and Schneider (2012), focussing on the scientific genre.

1.4.2 Significance Testing on Noisy Data

I have conducted significance tests on each of the genre difference figures presented in the previous section. While statistically the differences are highly significant, we need to bear in mind that the parser has a certain error rate. It is theoretically possible that the differences between the genres are largely due to systematic parser performance

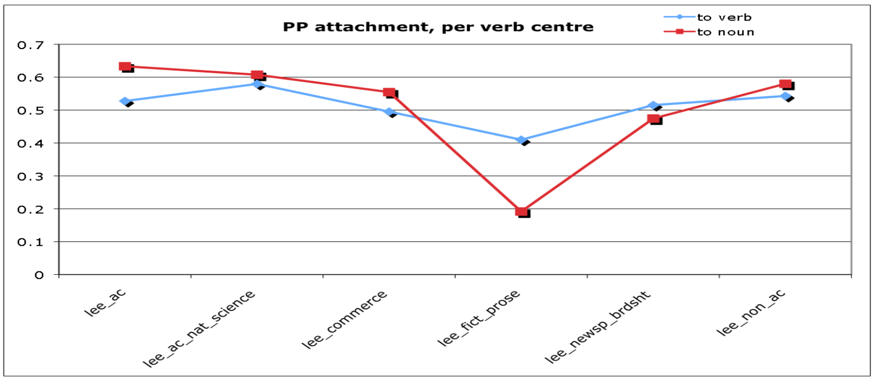


Figure 1.3: PP-attachment per verb-centre

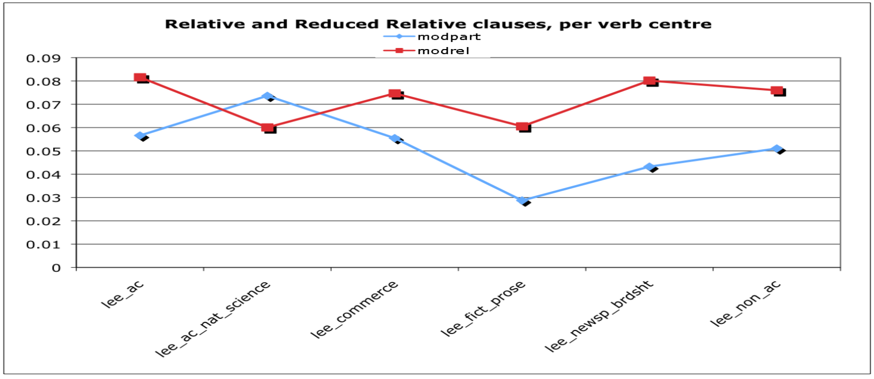


Figure 1.4: Relative and reduced relative clauses per verb-centre

differences in different genres. If one can show that the parser performance does not depend on the genre, that the errors (noise) are nearly randomly distributed across genres, then the remaining errors simply appear as so-called 'white noise' which does not add a systematic skew. In the worst case, if the random noise is very strong, it may be so strong that a significant difference is not detected as significant (type II error), while the risk for wrongly postulating a significant difference (type I error) is low. In other words, we most likely err on the side of caution if we can show that the performance is similar across the investigated genres.

In order to minimise the risk of an influence of parser performance, I have evaluated the performance of the parser on the standardized 500-sentence evaluation corpus GREVAL (Carroll, Minnen, and Briscoe, 2003), 100 random sentences from the BNC, 100 random sentences from the biomedical science corpus GENIA (Kim et al., 2003). I have manually annotated the random sentences from BNC and GENIA. The performance given in Table 1.7 shows that about 90% of all reported subjects and objects are correct (precision), and over 80% of all subjects and objects in the text are found (recall), and that performance on the highly ambiguous PP-attachment relations $modpp$ and $pobj$ is easily 10% lower. It also shows that while performance fluctuates considerably, there is no trend to systematically lower performance in different genres. A more detailed evaluation on biomedical texts is given in Haverinen et al. (2008). Detailed evaluations using GREVAL are also given in (Schneider, 2008).

I have also conducted a χ^2 contingency test on the performance evaluation data. The probability does not reach any significance level, and thus does not allow the rejection of the null hypothesis (i.e. the hypothesis that there is no significant difference), so no systematic difference in performance between the evaluated corpus subsets can be claimed. Such a test does not permit the rejection of the reverse claim (that parser error rates on the tested relations are spread homogeneously over genres). But our suspicion that parsing mistakes could have a major

Newspapers (GREVAL)	Percentages for some relations, 500 sentences			
	Subject	Object	modpp	pobj
Precision	91	89	73	74
Recall	81	83	67	83
Scientific (GENIA)	Percentages for some relations, 100 random sentences			
	Subject	Object	modpp	pobj
Precision	90	94	83	82
Recall	86	95	82	84
BNC written	Percentages for some relations, 100 random sentences			
	Subject	Object	modpp	pobj
Precision	86	87		89
Recall	83	88		70

Table 1.7: Performance on newspaper (GREVAL), biomed. science (GENIA) texts, and BNC written

ICE-GB	A posteriori checking on 100 random sentences			
	Subject	Object	modpp	pobj
Precision	97	92	82	90
Recall	92	88	79	92
ICE-FIJI	A posteriori checking on 100 random sentences			
	Subject	Object	modpp	pobj
Precision	99	94	83	72
Recall	87	100	91	76

Table 1.8: *A posteriori* Performance on ICE-GB and ICE-FIJI

influence does not materialise.

In the following sections, I will summarise the use of a parser-based approach to the detection and description of regional variation. Again, the same potential criticism needs to be addressed: there could be a significant relation between regional variety and parser performance. We have investigated if this is the case in Schneider and Hundt (2009). Our initial assumption was that parser errors and parse fragmentation may be used to detect regional characteristics. This was not the case, which could be seen as disappointing news. As good news it emerged, however, that error rates are similar across regional varieties, which means that for many phenomena the parser can be used reliably. Table 1.8 shows the comparison of ICE-GB (a central L1 variety) to ICE-FIJI (an undisputed L2 variety), on 100 random sentences. No systematic performance difference can be seen, and the data is not significantly different between the two varieties, according to χ^2 contingency test. Again, there is only little variation, which indicates a large white noise proportion. In particular, the probability of *chance difference* **in relation frequencies** *between genres*, according to χ^2 testing, is several orders of magnitude less likely than *chance difference* **in performance** *between genres*. The evaluation method that I have used for this experiment (Table 1.8) was *a posteriori* checking, which means that the parser output was checked. In the evaluations shown in Table 1.7 on genre variation I have annotated the random sentences from scratch. As one tends to be more lenient towards accepting an analysis, the performance reported here cannot be directly compared to previous evaluations, and is only suitable for the purposes of comparison rather than as a reliable measure of the general performance of the parser.

1.4.3 Data-Driven Case Study of Regional Variation: Indian English

I give a case-study example for a data-driven approach to the detection of regional variation using a parser-based approach. I have investigated Indian English as test case in Schneider (2013).

Regional differences are often more subtle and intricate than genre variation. This is also a major reason why our initial hypothesis in Schneider and Hundt (2009), that parser errors and parse fragmentation could be used to detect regional characteristics, had to fail. Most sentences in ICE-India could also have been produced by a Native British or American speaker, there is nothing ‘unusual’ in them. Schneider (2004) has observed that these subtle differences typically occur at the level of lexicogrammar rather than at the level of syntax.

> [D]istinctive phenomena tend to concentrate at the interface between grammar and lexicon, concerning structural preferences of certain words (like the **complementation patterns that verbs** allow), **co-occurrence and collocational** tendencies of words in phrases, and also patterns of word formation. (Schneider, 2004, 229, boldface added).

Concerning verbal complementation patterns, I have investigated (1) verb formation: the ditransitive complementation is particularly restricted and depends on verb semantics.

Concerning co-occurrence and collocational tendencies, I have (2) detected collocations by using statistical distribution measures such as mutual information, Z-score or Observed/Expected (O/E). I compare O/E from ICE-India divided by O/E from British English (BNC).

I have also used a simple surface approach (3) testing which frequent ICE-India trigrams are absent in the BNC. The initial lists were dominated by proper names, which I thus had to filter by discarding all trigrams containing proper name tags. I elaborate on these three steps in the following.

(1) Ditransitive Complementation When comparing the frequencies of verbs heading double-object constructions between ICE-GB and ICE-India (Table 3b in Schneider (2013)), the lists largely overlap, the verbs *provide, grant, develop, hand* are either absent in ICE-GB or considerably less frequent. Such tendencies are well known and have been reported, e.g. in Mukherjee and Hoffmann (2006).

An example of the 'new' ditransitive verb *provide* is:

(15) I am enclosing herewith a detailed resume of my professional career and feel that I can *provide you the best possible services* in the areas required. (ICE-India W1b-024)

(2) Collocations that are considerably more frequent in ICE-India I have compared trigram frequencies between ICE-India and British English as follows. I have calculated O/E for each trigram in ICE-India, taking the BNC value as expected value per 100 million words for British English. Initially, I used ICE-GB for British English, but data sparseness turned out to be a serious problem. I obtained better results when using the larger BNC corpus.

The majority of the hits which I thus obtained (Table 2 in Schneider (2013)) arise from text selection criteria, for example there are relatively many legal texts in ICE-India (*proviso to section*, *statement before the*), many medical texts (*the blood group*), and the spoken data percentage is much larger, showing hesitations etc. (*a very very*, *in the in*). I also found quite formal expressions (*do not recollect*) and, as it turns out, zero articles (*for number of* and *on right side*), i.e. expressions involving an NP where British or American English speakers would use an article, but Indian English speakers often do not use any. An example of the zero article is:

(16) And *for number of* years following the Nehruvian outlook this society has built itself. (ICE-India S1b-054)

After thus having identified that zero articles are a potential feature of Indian English, I have tested a large subset, consisting of two thirds of the written part of several ICE corpora. In ICE-GB, 10,034 of the 27,360 singular common nouns, or 36.7%, have no article. In ICE-India, 12,633 of the 29,032 singular common nouns, or 43.5% have no article. The difference is statistically highly significant (chi-square contingency test, $p < 0.01\%$).

I have investigated zero articles by genre. While the percentage is spread homogeneously across genres in ICE-GB, ICE-India shows a peak in the least edited genre, student essays, and a tendency towards over-correction in the most edited genre, press, as Figure 1.5 shows.

Zero articles are recognised as being essential (Sand, 2004, 295). It is very difficult to measure zero-forms in a surface-based approach, however (e.g. (Sedlatschek, 2009, 198)). In a syntactic approach, a zero (article) form is simply a base noun phrase (=noun chunk) without article; measured per noun. In order to approximate the envelope, I only measure singular non-proper names noun chunks (POS tag $\_NN$), as plural indefinite and proper names are typically without article. The syntactic approach thus allows us, on the one hand to measure zero-forms, and on the other hand to approximate the envelope of variation.

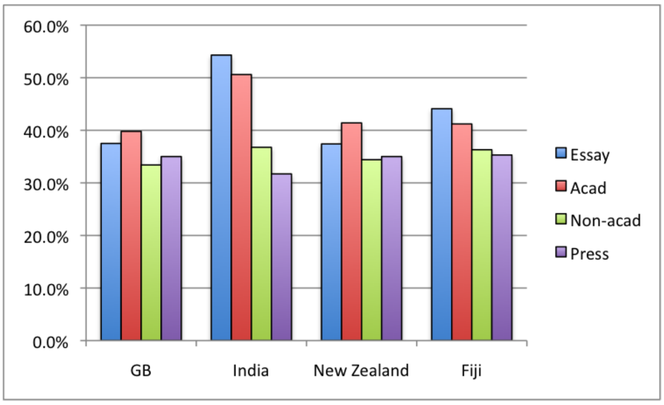


Figure 1.5: Zero-form article percentages per *_NN*-tagged head noun (singular common noun) across genre and variety

O/E ratio	Head	Prep	f(India)	O/E (India)	O/E(BNC)	manual inspection comment
80.6962	discuss	about	10	148.012	1.83	You come we will *discuss about* it.
51.3664	study	about	7	67.71	1.31	Today we are *studying about* rotation and revolution of the earth.
705.33	advise	into	7	279.73	0.39	no, consistent parsing error
39.8306	result	into	5	55.36	1.39	This *resulted into* a deep sense of growing loneliness
78.7867	burst	of	5	234.214	2.97	no
53.0517	arrest	from	5	59.37	1.12	five more terrorists *were arrested from* his home
93.5978	etch	at	3	147.23	1.57	no
67.2343	withstand	to	2	139.35	2.07	no
46.6381	significant	on	2	33.16	0.71	no
45.8399	nice	on	2	70.01	1.52	no
84.4974	line	of	2	120.45	1.42	no
47.4123	land	into	2	102.12	2.15	Atul's tendency of worrying too much ... *landed him into* trouble
107.968	exciting	on	2	315.06	2.91	no
214.685	benefit	out	2	128.15	0.59	yes: So they*'ll benefit out of the faculty teaching*

Table 1.9: Verb-preposition collocations in ICE-India

(3) Trigrams from ICE-India that are absent in BNC While some trigrams are considerably less frequent in the much larger BNC than in ICE-India, there are also some trigrams from ICE-India that are absent in BNC (Table 3a in Schneider (2013)). Besides sampling issues, the list shows Indian features like archaic spellings (*now a days*), formal language (*the honourable minister*), unusual verb complementation with prepositional phrases (*is called as*) appear in the list. An example for *is called as* is:

(17) A substance which is helping in chemical reaction *is called as* a reagent. (ICE-India S1b-004)

A number of L2 Englishes including Indian have been described as using novel verb-PP combinations like *is called as*, e.g. (Sedlatschek, 2009) reports. I use a data-driven method to investigate them. I compare collocation measures by dividing the O/E values of ICE-India by those from BNC. Thus collocations that have high collocation strength in Indian English but not in British English should get high scores.

$$O/E\ ratio = \frac{O/E(India)}{O/E(BNC)} = \frac{\frac{O(India)}{E(India)}}{\frac{O(BNC)}{E(BNC)}} = \frac{\frac{O_{India}(R,w_1,w_2) \cdot N_{India}}{O_{India}(R,w_1) \cdot O_{India}(R,w_2)}}{\frac{O_{BNC}(R,w_1,w_2) \cdot N_{BNC}}{O_{BNC}(R,w_1) \cdot O_{BNC}(R,w_2)}} \tag{1.5}$$

where N is corpus size, R is the relation (verb-attached PP = $pobj$), w_1 the head (verb), w_2 the preposition. I have also tested T-score as collocation, and obtained slightly worse results. A sample result is given in Table 1.9. The last column contains my manual inspection of the results, showing an example or *no* if it is a false positive.

#	mcorp	beheaded chunk:				
		word_tag sequence	O	E	Freq.	potential TAM cat.
2	37.18	are_vbp	189.08	151.90	421	progressive or passive?
3	25.61	should_md	70.96	45.35	158	modality
4	23.43	will_md	150.45	127.02	335	future
5	20.13	should_md be_vb	47.16	27.03	105	modality
6	13.12	were_vbd	133.84	120.72	298	past progressive or passive?
7	11.15	can_md	92.07	80.92	205	modality
8	10.82	was_vbd	196.71	185.89	438	past progressive or passive?
9	7.15	have_vbp	101.50	94.35	226	perfect
10	7.03	could_md be_vb	25.60	18.57	57	modality
11	6.49	are_vbp not_rb	17.52	11.03	39	progressive or passive?
12	5.87	will_md be_vb	36.38	30.51	81	future
13	5.01	need_vbp to_to	11.23	6.22	25	modality
15	4.19	would_md be_vb	24.25	20.06	54	modality
16	4.17	are_vbp being_vbg	8.98	4.81	20	progressive
17	3.88	needs_vbz to_to	6.29	2.40	14	modality
19	3.44	have_vbp to_to	13.47	10.03	30	modality
20	3.39	should_md not_rb	8.53	5.14	19	modality

Table 1.10: TAM profile for ICE-Fiji

An important conclusion that can be drawn from my research in Schneider (2013) is that even though language models such as syntactic parsers are imperfect and produce a certain level of errors, they can pick up a signal, even on out-of-domain texts, even without modelling an envelope of variation, and without even coming close to reaching statistical significance. As we see in the fourth column (*f(India)*) frequencies are very low, but even hapax legomena have delivered many true positives.

I have used the same approach for the detection of novel verb-preposition constructions ICE-Fiji in Schneider and Zipp (2013). Experimentally, I have used larger web corpora, which partly leads to better results as data is less sparse, but partly also to worse results as corpora collected from the web are more skewed and contain duplicates.

1.4.4 Tense, Aspect and Modality in the ICE corpora

Schneider and Hundt (2012) present an even more radically data-driven approach, by looking at the verb chunks across a large selection of ICE corpora (we have used GB, New Zealand, India, Fiji and Ghana). After discarding the lexical part of the verb chunk (the head) what remains is a morphosyntactic trunk (the beheaded chunk) containing auxiliaries pointing to tense and aspect, modals expressing modality, and adverbs. If one compares the frequencies of these trunks across several ICE corpora one gets a tense-aspect-modality (TAM) profile for each variety.

For Mair (2009, 18), modals of obligation and necessity are “almost perfect diagnostic to assess the synchronic regional orientation of a New English with regard to British or American norms and also its degree of linguistic conservatism.” In our approach, I calculate a measure m as the difference of O and E. Here, E is the random distribution of beheaded verb chunks across the 5 corpora. For each beheaded chunk from e.g. ICE Fiji m is:

$$m = f(fiji) - \frac{f(fiji) + f(India) + f(GB) + f(NZ) + f(Ghana)}{5} \qquad (1.6)$$

The profiles that emerge largely confirm existing observations, but they often demand very careful interpretation. A shortened sample profile for ICE-Fiji is given in Table 1.10 for illustration, our interpretation is given in the last column. The list shows for example (lines 2,6,8,11) that present progressives are used very frequently in ICE-Fiji, and that the modal verbs *will, would, need, should* (lines 3,4,5,12,13,15,17,20) are used more often in ICE Fiji. The beheaded chunk in lines 2,6,8,11 could equally point to passives, I have summed over the head tags to see that progressive forms are responsible. The complete absence of an envelope of variation also entails that random or document-specific characteristics can lead to inflated numbers and skew the lists. For example, 3rd person singular

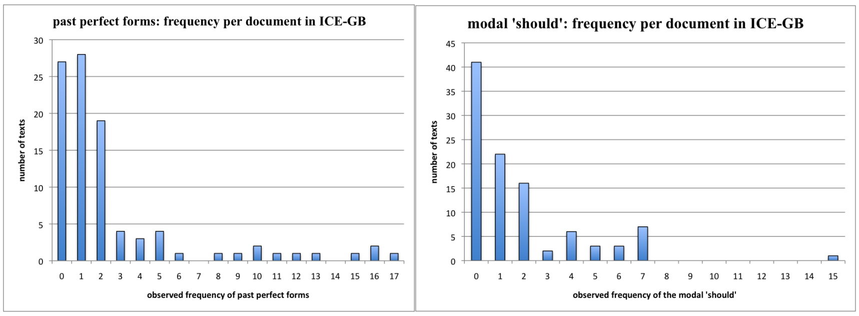


Figure 1.6: Frequencies of past perfect forms and of the modal verb *should* in ICE-GB

SEM	s-gen #	% s/all	of #	% of / all
B-Brown-J (1930)	240	9.8%	2207	90.2%
Brown-J (1960)	319	15.5%	1737	84.5%
Frown-J (1990)	564	25.6%	1637	74.4%
1930-60		+32.7%		-21.3%
1960-90		+76.8%		-5.8%

Table 1.11: Frequency and percentages in the envelope, according to the automatic SEM method

verbs are considerably more frequent in ICE India, and plurals less frequent, for no apparent reason. Checking per-document counts also indicated that significance testing may be unreliable. For example, Figure 1.6 shows frequencies of past perfect forms and of the modal verb *should* in ICE-GB. If there distribution were independent of the document, which significance tests typically assume, one would see a Gaussian bell-shaped distribution. We will come back to this issue in chapter 3.

1.5 Language Change

1.5.1 Development of the Saxon Genitive

Röthlisberger and Schneider (2013) investigate the development of the Genitive alternation from 1930 to 1960 to 1990 using the scientific section of the Brown & LOB corpus family. We have seen in section 1.3.2 that the envelope of variation (Labov, 1969; Sankoff, 1988) needs to be approximated in order to model the speaker's choice. I have presented an approach using lexical overlap in that section. We compare three methods (RAW, MAN and SEM) approximating the envelope of variation in the Saxon Genitive alternation: in RAW, no filter is used, I simply count all Saxon Genitive occurrences and all of-PPs. This approach serves as a baseline. In MAN, we have manually examined each token and decided if it can be in the alternation, using agreed criteria. This gives us the gold standard. In SEM, I use the lexical overlap method, as described in section 1.3.2. Due to data sparseness, I had to use semantic class overlap, according to WordNet.

The frequency of the Saxon Genitive is increasing highly significantly over time. Table 1.11 shows the percentages in the envelope, according to the SEM method. The percentages according to the three methods are compared in Figure 1.7. The graph shows that SEM (the lexical overlap method described in section 1.3.2) is a better approximation to MAN than RAW. The table also illustrates again that the use of the Saxon Genitive increases. Another insight that we have gained in Röthlisberger and Schneider (2013) is that the importance of the principle of end-weight has increased over time in the Saxon Genitive alternation.

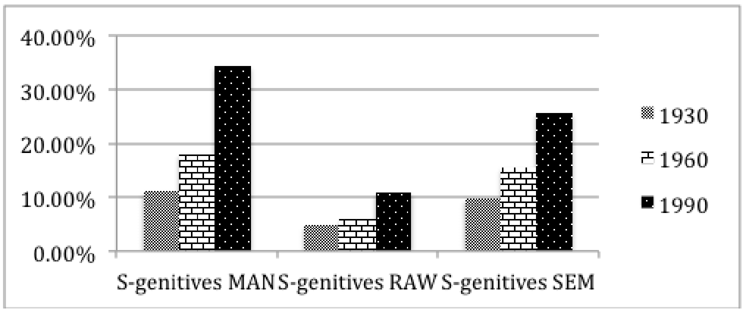


Figure 1.7: Percentages of Saxon Genitives in the envelope

1.5.2 Adapting the Parser to Earlier Diachronic Stages

I have also used the parser to investigate stages of the English language that are earlier than the 20th century. A major problem is that parser performance gradually declines when one moves further back in time. In Schneider (2012a) and in Schneider, Lehmann, and Schneider (2014) I have evaluated the performance of the parser on the ARCHER corpus (Biber, Finegan, and Atkinson, 1994), a genre-balanced corpus containing texts from 1600 to 1999, from Early Modern English (EModE) to Late Modern English (LModE) to Present-Day English (PDE). The most obvious source of tagging and parsing errors was caused by the fact that spelling in the earlier Archer texts was different from PDE and not fully standardized. Applying standard taggers to them produces many tagging errors, many of which are due to spelling differences. A solution to this problem can be to map the original spelling to its PDE counterpart, for which a number of tools are available. We have used VARD (Baron and Rayson, 2008).

Tagging and consequently also chunking and parsing, profit from mapping the original spelling to the same spelling as used in the tagger and parser training resource. The statistical performance disambiguation, which uses lexical heads, can equally profit. As the normalisation process also makes errors, the assumption that performance will improve cannot be taken for granted.

Concerning tagging accuracy, this assumption has been tested in Rayson et al. (2007). They report an increase of about 3% (from 82% to 85% accuracy) on Shakespeare texts. Concerning parsing, the assumption that normalisation improves the performance has, to our knowledge, not been confirmed before Schneider (2012a). I have thus tested it by selecting 100 random sentences from the 17th century (1600 – 1699) part of the Archer corpus, and found better syntactic analysis due to VARD in 12 sentences, worse syntactic analysis due to VARD in 1 sentence, and improvements paralleled by new errors in 3 sentences. In Schneider, Lehmann, and Schneider (2014) I have used a 422 sentences random set from the ZEN corpus and found 68 improved and 5 worse analyses.

An example sentence from an early ARCHER text, is given in 18.

(18) The ship, the Amerantha, had never yett bin att sea, and therfore the more daungerous to adventure in her first voyage; butt she was well built, a fayre ship, of a good burden, and had mounted in her forty pieces of brasse cannon, two of them demy cannon, and she was well manned, and of good force and strength for warre: she was a good sayler, and would turne and tacke about well; she held 100 persons of Whitelocke's followers, and most of his baggage, besides her own marriners, about 200. (ARCHER 1654whit.j2b)

I have conducted detailed evaluations on 100 sentences in Schneider (2012a) and on 400 sentences in Schneider, Lehmann, and Schneider (2014), both from the ARCHER corpus. F-score performance on the 400 sentence random set is given in Figure 1.8 on the left. As expected, parser performance decreases for the 18th and 19th centuries, and shows a steeper decline for the texts before 1700. F-score is the harmonic mean of precision and recall. Some errors were easy to correct, for example *but* as adverb as in sentence 19 is not known in the parser grammar.

(19) He is such an Itinerant, to speak that I have *but* little of his company. (Archer:1766aadm)

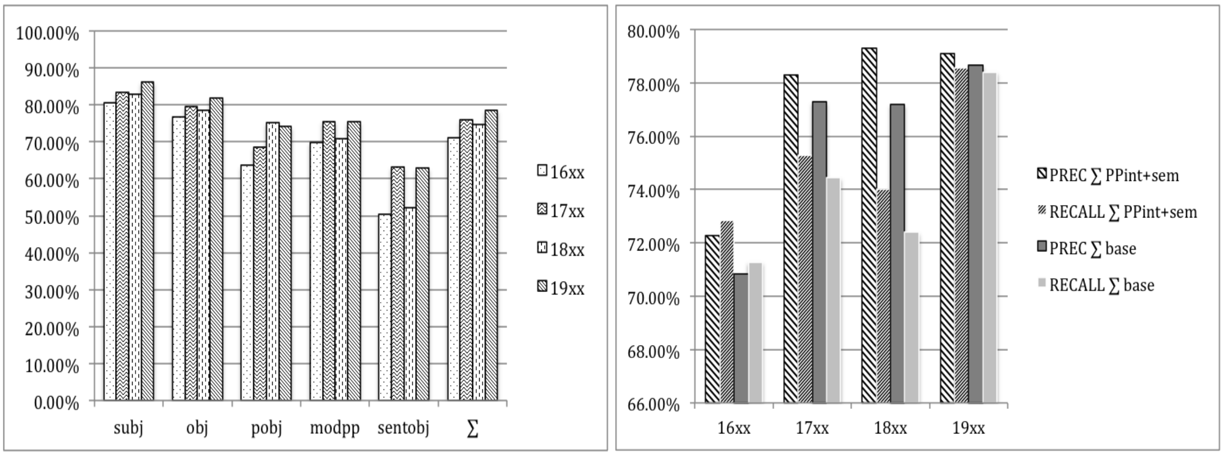


Figure 1.8: F-Score performance of the ARCHER baseline system, by century and syntactic relation (on the left), and (on the right) F-Score of the ARCHER baseline system (base) compared to the improved system (int+sem)

Additional parsing errors in the texts before 1900 come from a number of sources, including rare 'poetic' constructions that are not licensed by the grammar, preferences that do not match (examples include the genitive of quality *ship of a good burden* in example 18), and high complexity, marked constituent order. In this category, we find parser errors that also occur in PDE, but they are more frequent in ModE. Sentence complexity and length is generaly higher in historical texts (Fries, 2010, 31).

Changing the grammar rules typically has the following effect: if one constrains rules too much, the correct reading can often not be found, for example if a marked constituent order is used. If one constrains too little, ambiguity explodes, and the risk for incorrect disambiguation increases. Disambiguation can often be improved by adding more resources. One way to help disambiguation is to include more semantics and context.

Concerning semantics, I have included semantic expectations, which I will explain in section 2.5.5. Concerning context, I have added a PP-interaction model that I have described in Schneider (2012b). It is well known that considering sister, grandmother and great-grandmother nodes increases parsing accuracy (e.g. (Charniak, 2000), (Bod, Scha, and Sima'an, 2003)), particularly in the case of the highly ambiguous PP-attachment relations.

These two measures improve recall and precision, as is summarised in Figure 1.8 on the right. The performance of the baseline parser is shown by grey bars, and the performance of the extended parser by striped bars. We observe two interesting facts: first, earlier centuries profit more from the adaptation, which I believe may indicate that, due to freer word order and longer sentences, constraints on semantics and complexity are more important. This provides an example of how a computational linguistic tool, which I aimed to improve for PDE texts, by means of its application to different research areas (the parsing of Late Modern English texts), leads to improved results.

I have also tried additional adaptations to the historical texts, but could not further improve the performance. We conclude that what helps most is spelling normalisation and resources using semantics and context that were developed to improve the PDE system. In this sense, texts from about 1700 on are indeed similar to PDE but harder. López-Couso, Aarts, and Méndez-Naya (2012) state that the Late ModE period is marked by regulatory and statistical changes, while there are only few grammatical innovations, namely the progressive passive and the *get*-passive. The statistical changes typically mentioned are that the progressive form increases in frequency, that *be* as auxiliary decreases, that periphrastic *do* is fully established, and that non-finite complementation and relativisation have undergone changes. We have investigated relativisation in more detail in Hundt, Denison, and Schneider (2012) and Hundt, Dension, and Schneider (2012). While sentence length decreases over time, information compression (Biber, 2003; Leech et al., 2009) increases equally radically: the number of full relative clauses decreases, participial clauses (sometimes called reduced relative clauses) increase, and become the prevalent form of postmodification in the scientific genre, as we have seen in Figure 1.4. Also other forms of compressed noun modification increase.

1.6 Intermediate Conclusions

In this chapter, I have mainly used approaches which measure the signal that an automatic parser emits. I have assessed the quality of the signal by measuring the parser performance and the parser noise. We have seen a key insight of Schneider and Hundt (2009) in section 1.4.2:

- Concerning varieties and genre, parsing errors largely lead to white noise, homogeneously distributed errors, which do not add a consistent skew to our data.
- The signal (correct analyses) is stronger than the noise (parsing errors).
- The use of a syntactic parser as a research method is suggested.

Noise levels (error rates) are considerably higher on historical texts, but the error rate can be lowered by adding more semantic and contextual models. The quality of tagging will likely further improve if more tagged texts from the period will be available, from which parsing will equally profit.

In section 1.2, I have highlighted some of the key findings of Lehmann and Schneider (2009). The key findings are:

- Performance of my parser (Schneider, 2008) is not considerably lower on BNC written texts than on the training domain of newspaper texts.
- There are strong collocational restrictions in the subject-verb and the verb-object relations, confirming Hoey (2005)'s hypothesis that lexical priming occurs everywhere.
- Parsed data delivers considerably cleaner results (Seretan and Wehrli, 2006; Seretan, 2011), so that the O/E collocation metric can be used profitably, despite its tendency to overreport rare collocations.

I have also illustrated some of the findings of Lehmann and Schneider (2011). The full list of key findings of Lehmann and Schneider (2011) include:

- A confirmation of Pawley and Syder (1983, 193) "that native speakers do not exercise the creative potential of syntactic rules to anything like their full extent".
- Frequency-based measures work well as a stopgap to semantic non-compositionality (Wulff, 2008). We will see in chapter 3 that this applies to many semantic tasks.
- A realistic approach to construction grammar (Stefanowitsch and Gries, 2003) requires large amounts of data. I have applied collocation measures to corpora of 1 to 10 millions of words in Schneider (2013) and Ronan and Schneider (submitted) and obtained worse results than on the 100 million words BNC.
- Yule's K (Yule, 1944) as a measure of fixedness improves the detection of idiomatic structures.
- Idioms and collocations are a gradient phenomenon.

In section 1.3 I have shown that even the passive alternation is restricted (Lehmann and Schneider, 2009) and illustrated key insights of our exploration of the dative shift in Lehmann and Schneider (2012b):

- We provide detailed gradient list and descriptions of lexical preferences in the dative shift.
- We attain improved results due to approximating the envelope of variation with our lexical overlap method.
- We have seen that the dative shift is not binary and an open list.

In section 1.4 I have illustrated genre variation and summarised a corpus-driven (also called data-driven) case study of Indian English (Schneider, 2013) allowing one to detect:

- new double object ditransive verbs

- zero-form article uses
- new verb-preposition combinations

We have also seen that even if no significance can be reached due to sparse data, combining automatic methods and manual sifting leads to new insights.

For my research, I have used data-driven methods, most radically so in Schneider and Hundt (2012), in section 1.4.4, where we have learnt:

- When the envelope of application is not constrained, data-driven approaches need to be interpreted carefully. Increased frequencies may have many causes: for example, high frequency of the participle tag (VBN) may both indicate use of passive or use perfect aspect.
- High frequency of a certain modal verb may indicate generally higher use of modals (as is the case in e.g. ICE-EA).
- Significance tests may report significant differences between side-effects. We will return to this problem in chapter 3, where I show that models allow one to take a multitude of factors and their interactions into consideration.

The insistence on lexical interactions and data-driven approaches has ‘ugly’ consequences. It leads to enormous amounts of idiosyncratic data. On the one hand complexity is the characteristic if not quality of linguistic data. These long lists are an important result themselves, they can for example be used for dictionary construction, as has been done in the COBUILD dictionary by John Sinclair (e.g. (Moon, 2009)). On the other hand, the complexity of the data, which stems form a large, potentially infinite number of sources (or what one calls *factors* or *features* in statistics) is a main motivating factor for the use of models in chapters 2 and 3: only they can pay respect to the many interconnected factors that come into play. If not only the factors are interconnected, but also the outcomes to which these factors lead, then one needs more complex models. This is what I discuss in chapter 3.

Data-driven approaches are for example used to discover collocations (Evert, 2009), or diachronic word class shifts (Mair, Hundt, Leech & Smith, 2002). For the discovery of collocations, word forms or lemmas are used as uncontested features, for word-class shifts agreed-on part-of-speech tags can be used. I have used syntactic patterns as uncontested features for the detection of collocations (section 1.2), regional variation features like ditransitives and verb-PP constructions (section 1.4), and lexical overlap (section 1.3) for constraining the envelope of variation, based on the assumption that two lexical items in combination usually disambiguate each other’s word senses to a large enough degree to be used as equivalence class.

We need to discover, however, that in the case of alternations, there is a considerably less stable base than in collocations or part-of-speech tags, as Arppe et al. (2010) warn us. In particular, there are manifold restrictions, strong lexicogrammatical interactions, alternations are not binary, and the sheer number of alternation is contested. I will take up these points in chapter 2 again.

In section 1.5 I have applied the lexical overlap method to approximate the envelope of variation to the Saxon Genitive in Röthlisberger and Schneider (2013) and shown the following.

- The lexical overlap method leads to better results.
- The frequency of the Saxon Genitive has increased over time.
- The importance of the principle of end-weight has increased in the Saxon Genitive alternation.

I have adapted my parser to earlier diachronic stages. Key insights of Schneider (2012a) and Schneider, Lehmann, and Schneider (2014) are:

- Spelling normalisation with VARD (Baron and Rayson, 2008) leads to improved parsing.
- I have extended the grammar to address specific shortcomings.
- Adding more context information and including semantic expectations (see section 2.5.5) improves parsing of historical texts more than on PDE texts.
- The Late ModE period is marked by regulatory and statistical changes, while there are only few grammatical innovations.

Chapter 2

Linguistic Approaches in Biomedical Text Mining

Information Retrieval (IR) is concerned with the task of finding information in large text collections. Linguistic approaches can help the process. For example, including synonyms in a search, or the use of syntactic distances instead of surface word distances may improve recall. Biomedical texts are an important research and application area for IR. Here I report on our approaches to IR and Text Mining (a sub-branch of IR) in the biomedical area, how the use of a parser improves IR, and how, in turn, linguistic insights have been gained.

2.1 Motivation

The efficient localisation of specific information in large document collections is an important research task. For example, in the area of biomedical science, the PubMed repository contains more than 20 million publications (`http://pubmed.gov`). Keeping track of the soaring number of relevant research results is increasingly difficult, even for experts. Automatic tools and shared resources which offer support in this knowledge management task are vital. Scientists need to find relevant articles and know about the central results and approaches. This knowledge management task includes, among others, the detection of the following relations (also called *interactions*) between entities:

- protein-protein interactions: e.g the BioCreative research challenges (Hirschman et al., 2005; Krallinger et al., 2008; Leitner et al., 2010; Krallinger et al., 2011) and the BioNLP challenges
- drug-disease interactions: e.g. PharmGKB (Sangkuhl et al., 2008) and CTD (Wiegers et al., 2009). I discuss an approach to this task in sections 2.2 ff.
- methods and approaches used by the authors of a scientific publication: e.g BioCreative III research challenge PPI-IMT (Krallinger et al., 2011), which I discuss in section 2.5.3

IR approaches can and have been used to address these tasks. A classical IR application is for example a search on the web for individual words or surface strings. Text Mining (see e.g. Cohen and Hunter (2008) for a brief introduction) goes beyond word-based IR approaches by using more high-level patterns, such as integrating information on synonyms or syntactic parsing, co-occurrence windows, rules, and machine learning. Relation Mining is a subdiscipline of Text Mining which specifically searches for qualified relations between the domain-specific entities of interest. For example, in genetic biomedical research, *how do a protein and a gene interact?*; in pharmacological research, *which side-effects can a drug have?*; in the political domain, *what does a party think about a certain topic?* (e.g. Schneider (accepted for publication),Wueest, Schneider, and Amsler (2014)). Results obtained by Relation Mining can be stored as semantic web triples (see e.g. Schneider and Zimmermann (2010)). They closely resemble natural language *subject-verb-object* or thematic *agent-verb-patient* relations. The BioNLP challenges explicitly use the thematic roles (in the sense of Fillmore (1968)) *agent* and *target* and typed relations, some BioCreative tasks use typed actions (so-called action terms).

Our techniques have been validated by participation in several Text Mining competitive evaluation challenges (BioCreative II (Rinaldi et al., 2008), BioCreative ii.5 (Rinaldi et al., 2010b), BioCreative III (Schneider, Clematide, and Rinaldi, 2011; Rinaldi et al., 2010a), BioNLP 2009 (Kaljurand, Schneider, and Rinaldi, 2009), BioNLP 2011 (Tuggener et al., 2011), CALBC (Rinaldi, Clematide, and Schneider, 2010)). The results of our group were among the best reported in several of these challenges. For example, in BioCreative ii.5 (2009) we achieved the best results in finding mentions of protein-protein interactions. In BioCreative III (2010) we were one of only two groups which participated in all tasks, always achieving competitive results. I will only present a part of the research of our team in this chapter: those studies in which the author was centrally involved, and only those which are linguistically particularly relevant. We will see in the following that important linguistic insights can be gained in our Text Mining approach.

- syntax: we use a syntax-based approach to Text Mining. I summarise the approach in section 2.3
- alternations: I illustrate that alternation and research and Text Mining share a common core: to detect and map all semantically equivalent alternatives expressing the same thematic relation. I present the suggestion in section 2.4
- discourse: particular attention to discourse features is paid. I present our approach in section 2.5

2.2 A Pharmacogenomic Sample Task

As example of the listed Text Mining tasks we discuss a linguistic approach to drug-disease interactions. Pharmacogenomics and toxicogenomics study the relationships between drugs or chemicals, genes, and diseases, in particular in relation to specific individual mutations, which can affect the reactions to drugs and the susceptibility to diseases.

Important databases which provide a reference repository for such information are PharmGKB (Sangkuhl et al., 2008) and CTD (Wiegers et al., 2009). The information contained in PharmGKB and CTD is obtained through a combination of submitted experimental results and manual literature reading (so-called curation). The PharmGKB and CTD databases can be seen as large corpora with lean, document-level annotation. For many of the evaluation challenges, sentence or entity-level annotation is provided. We combine syntactic analysis and discourse features. I explain the syntactic features in section 2.3 and the discourse features in section 2.5, and show how much syntax and discourse features can improve Text Mining in terms of result ranking for interaction detection. We have used both an MLE and a Maximum Entropy approach.

2.3 The Syntactic Approach

Approaches to the identification of entity interactions based on syntax are now quite common; e.g. Rebholz-Schuhmann et al. (2006). There have been numerous publications showing the potential of dependency-based language analysis for Text Mining (e.g. Clegg and Shepherd (2007), Fundel, Küffner, and Zimmer (2007)). Pyysalo et al. (2007) describes a manually annotated corpus which includes a dependency based analysis of each sentence. Clegg and Shepherd (2007) use dependency graphs in order to evaluate four publicly available natural-language parsers.

Simple approaches apply handcrafted rules, e.g. regular expressions for surface searches (Giuliano, Lavelli, and Romano, 2006), or syntactic patterns on automatically parsed corpora (Rinaldi et al., 2006; Fundel, Küffner, and Zimmer, 2007). These approaches typically achieve high precision at the cost of recall. Rinaldi et al. (2006) detect subject-verb-object patterns for a manually collected list of verbs and applies a closed set of linguistic alternations like nominalisations and the passive voices alternation to extend coverage. I will explain this approach in section 2.4.

Syntactic approaches can be further enhanced by using machine learning methods, by extracting meaningful features from the dependency parse trees and from other intermediate stages of processing (e.g. Erkan, Ozgur, and Radev (2007), Kim, Yoon, and Yang (2008)). We have used semi-automatic approaches to the learning of useful syntactic configuration from a training corpus. The approach which I have developed in Schneider, Kaljurand, and Rinaldi (2009) and adapted to several evaluation challenges is summarised in the following. Like every typical

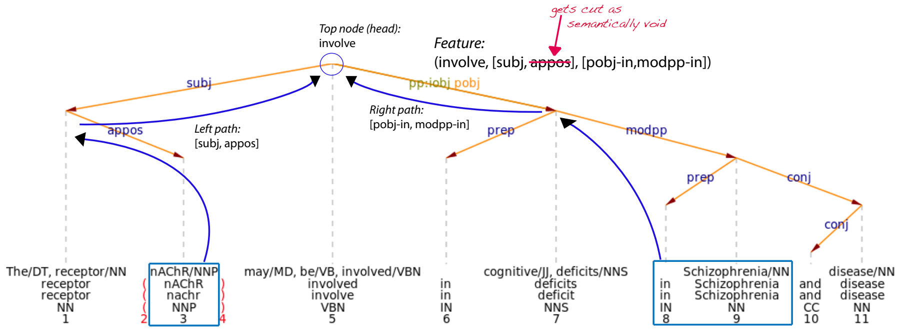

gets cut as
semantically void
Top node (head):
involve
Feature:
(involve, [subj, ~~appos~~], [pobj-in,modpp-in])
subj
pp:iobj pobj
appos
Left path:
[subj, appos]
Right path:
[pobj-in, modpp-in]
prep
modpp
prep
conj
conj
The/DT, receptor/NN receptor receptor NN 1
2 ((((
nAChR/NNP nAChR nachr NNP 3
)))) 4
may/MD, be/VB, involved/VBN involved involve VBN 5
in in IN 6
cognitive/JJ, deficits/NNS deficits deficit NNS 7
in in IN 8
Schizophrenia/NN Schizophrenia Schizophrenia NN 9
and and CC 10
disease/NN disease disease NN 11

Figure 2.1: Simplified internal syntactic representation, with illustrating notes, of the sentence “The **neuronal nicotinic acetylcholine receptor alpha7 (nAChR alpha7)** may be involved in cognitive deficits in **Schizophrenia** and Alzheimer’s disease.” from PubMed abstract 15695160

machine learning approach, we first learn rules and statistics from an annotated corpus in the training phase (section 2.3.1), then we apply them to unannotated data in the application phase (section 2.3.2).

2.3.1 Training Phase

We first parse the corpora (e.g. GENIA, PudMed parts, PharmGKB) with a dependency parser (Schneider, 2008). Entities are recognized and disambiguated using a pipeline based on LingPipe. For term recognition, we use a dictionary-based tool which delivers annotated document spans (terms), associated to a set of identifiers (concepts) from domain term databases. Our pipeline is described in more detail in Rinaldi et al. (2012). Here I focus on the detection of relations, also called interactions, between the detected terms.

All entities appearing in the same sentence are potentially interacting, thus the algorithm next records the syntactic path that connects them as a *candidate path*. The algorithm collects all candidate paths as described in the following. For each term co-ocurrence pair, i.e. two terms appearing in the same sentence, a collector traverses the syntax tree from each term up to lowest common mother node, and records all intervening nodes. Such traversals have been used in many biomedical interactions detection systems, e.g. Kim, Yoon, and Yang (2008). We will refer to the connection between the term pair which has thus been found as *path*. In Figure 2.1, such a path is shown blue.

If one records all the information that an intermediate node contains including its lexeme (e.g. *receptor, involve, in, deficit, in*) the path would be extremely specific, which would lead to sparse data and hence a recall problem for most applications. If one only records the grammatical role labels (e.g. *appos,subj,pobj,modpp*), the paths are too general, which leads to a precision problem for most applications. As an intermediate working solution, the algorithm records the lexical head lemma of the top node, and the grammatical labels plus prepositions connecting all intervening nodes. The algorithm splits the path into a left and a right half between the top node. An example of the path feature between *nAChR alpha7* and *Schizophrenia* can be seen in Figure 2.1.

The path feature is treated as a single feature. In many approaches path fragments are used. While this alleviates the problem that data is sparse, it neglects the fact that many semantic configurations are not local, they depend on considerably larger tree fragments, as we have learnt from the idiom principle in section 1.1. We use a single feature consisting of the entire path. To compensate the sparseness, I use very little lexical information (only the top node and the lowest node), and linguistic insights which allow us to shorten the typically sparse paths.

Prepositions are a closed class of lexical items that is crucial to syntax (Collins and Brooks, 1995; Collins, 2003), I have thus introduced them into the path.

Our paths are also shorter and less sparse than in many other representations. This is based on the linguistic insight that the syntactic paths that I use are shorter because they are based on chunks, and less sparse as I do not

#	p(relevant)	Head	Path1	Path2	TP	Count
1	13.62%	associate	subj	pobj-with	53	389
2	17.82%	associate	subj modpp-in	pobj-with	31	174
3	14.57%	effect	modpp-of	modpp-on	22	151
4	18.92%	effect	modpp-of	modpp-on modpp-of	21	111
5	20.65%	association	modpp-of	modpp-with	19	92
6	6.29%	be	obj modpp-of	subj	19	302
7	17.82%	metabolize	pobj-by	subj	18	101
8	29.63%	inhibit	pobj-by	subj	16	54
9	35.71%	associate	subj modpp-in	pobj-with modpp-of	15	42
10	23.81%	cause	subj modpp-in	obj	15	63
11	5.02%	be	subj	obj modpp-of	15	299
12	100.00%	analyze	subj modpp-in	pobj-in modpart pobj-with	14	14

Table 2.1: The most frequent path types in the PharmGKB training set, ranked by the number of true positives (TP)

include part-of-speech tags (which would fragment the data further).

In the next step the algorithm has to decide if a candidate path is a relevant path, in the sense that it expresses a relation between entities according to the training data. We have used several alternative approaches, I summarise three of them. As **first** approach, in the BioNLP challenges and the Biocreative protein-protein interaction challenges, where individual occurrences are marked as relevant in the training corpus, the distinction is obvious: only the candidate paths that are marked in the training corpus are relevant. In our sample sentence in Figure 2.1 there are $n = 3$ entities and hence $\frac{n(n-1)}{2} = 3$ paths: *nAChR alpha7* to *Schizophrenia*, *nAChR alpha7* to *Alzheimer's disease*, *Schizophrenia* to *Alzheimer's disease*. The first two paths are marked as relevant, the last one not. The path between *Schizophrenia* and *Alzheimer's disease* is $conj$, which only rarely expresses a relevant relation. In *between-PP* constructions it actually may: *the interaction between A and B*.

As **second** approach, in Schneider, Kaljurand, and Rinaldi (2009) we had no annotated training data for the relation detection task. I thus performed group-based manual annotation on GENIA, our learning corpus, and annotated the data myself. As we only had limited resources I performed a group-based annotation, starting with the most frequent path feature types. Eventually I annotated all 2500 paths which appeared at least twice. There were few path types in which different path instances suggested opposite decisions. We have learnt that for this task group-based annotation is a reasonable compromise, allowing one the annotation of relatively large amounts of data efficiently. A major advantage of annotating a large corpus over formulating hand-written patterns (section 2.4) is that no instance is missed (except for very rare ones that happen to be absent from a large corpus).

As **third** approach, in the CTD and PharmGKB data, each document comes with annotation of the IDs of relevant terms, but without evidence location. In such a lean annotation, sentence-level or entity-level is not provided. We thus used a weakly supervised, so-called distance-learning approach: if terms A and B are said to be in a relevant relation in a given document, then we assume that all syntactic connections between A and B in this document express a relevant relation. Such approaches are described in Rinaldi, Schneider, and Clematide (2012) and Buyko, Beisswanger, and Hahn (2012).

If the training resource states that a candidate path is relevant in one of the above senses, i.e. stating that both entities interact in the document, then we call it *relevant path*. The calculation of the number of *relevant paths* divided by the number of *candidate paths* gives us the Maximum-Likelihood probability that a path is relevant

$$p(relevant) = \frac{freq(relevant\ path)}{freq(candidate\ path)}$$

2.3.2 Application Phase

In the application phase, the system has to decide for each sentence in a corpus unseen in the training phase, if it contains a relevant path or not.

Ideally, one can apply $p(relevant)$ directly, but one typically encounters a serious sparse data problem when doing so. In Schneider, Kaljurand, and Rinaldi (2009) more than half of all paths are singletons, I have annotated the 2500 paths at least occurring twice. Accordingly, more than half of the paths encountered during the application phase can be expected to find no data from the training phase.

We address this acute sparse data problem in several ways. First, we try to strike a balance between expressiveness and generality in the design of our path feature, as we have described in section 2.3.1. Second, we use the following methods.

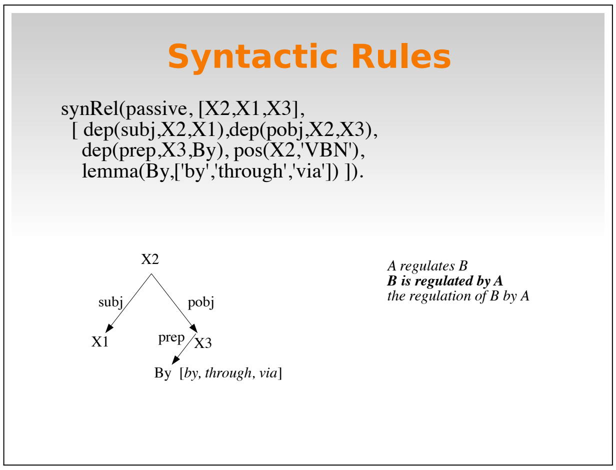
- using half-paths: we split the path between the the two entities in two halves: from the first entity to where the paths meet, and from the second entity to where the paths meet (the two rectangular brackets in Figure 2.1). This also allows us to profit from double argument relations for the detection of single argument relations, and vice versa.
- use additional features: we use many more features than the syntactic path: surface word sequences to recover from some parsing errors, pure occurrence (the most important entities are typically also the most frequent ones), document zoning (entities and concepts mentioned in the title or abstract are more important than those mentioned in the background section). An overview of the features used for one challenge is discussed in section 2.5.4.
- Maximum-Entropy classifier: Maximum-Entropy classifiers scale better, i.e. they perform better when large, complex, sparse data and many and partly related features are used. Maximum-Entropy is also known as statistical regression, which is increasingly used in linguistics to build models, as we discuss in chapter 3.
- Expand and filter abbreviations: we discuss in section 2.5 that acronyms can often be expanded to their correct full form in the document context. If acronyms and full forms can be mapped, this reduces the number of entity types in a document and thus alleviates sparseness.
- Transparent words (Meyers et al., 1998): I discuss transparent words in section 2.5. They have the effect that paths can be shortened, which alleviates sparseness.

2.4 From Relation Mining to Alternations

I have mentioned that some earlier Relation Mining systems, for example Rinaldi et al. (2006), formulate hand-crafted patterns and explicitly apply linguistic alternations to them. Figure 2.2 shows the passive verb rule. It says that a verb $X2$ which has a subject $X1$ and a passive agent $X2$ (a PP headed by *by, through* or *via*) is a *passive verb-agent-target* structure. Similar rules exist for active verbs and for nominalisations. The alternation rules which are given in Figure 2.3 map active transitive, passive, and nominalisation patterns. The alternation rule ensures that any event verb such as *regulate* will detect relevant structures in their various alternation forms, i.e. *A regulates B*, *B is regulated by A* and *the regulation of B by A*. Event verbs are typically domain-specific, the event verb rules are thus called domain rules. They may simply licence an event verb like *regulate* to instantiate the search in the application corpus for all occurrences of the event verb and its patterns as defined in the syntactic and alternation rules, or they can be used to describe more complex syntactic argument frames. Figure 2.4 shows the argument frame for the verb *trigger*. Such rule-based systems typically have good precision, but relatively low recall, as they miss many configurations; typically less direct ways of expressing relations, such as the relation between drug X and disease Y in

(20) Small doses of X have shown a positive effect on 23% of Y patients.

Machine learning systems which learn from annotated corpora are more data-driven (Tognini-Bonelli, 2001) in this respect. The weakly supervised learning phase on PharmGKB delivered the syntactic patterns shown in Figure 2.1 (and others, the figure is an excerpt). Line 4 shows the pattern which matches example 20. Comparing lines 1 and 5 shows an instance of the classical nominalisation alternation. Lines 2 and 9 show, however, that there are many other frequent ways involving the verb *associate* to express a relation – alternation options that our introspective approach of Rinaldi et al. (2006) missed. Also this list sharply tails off, more than half of the lines are

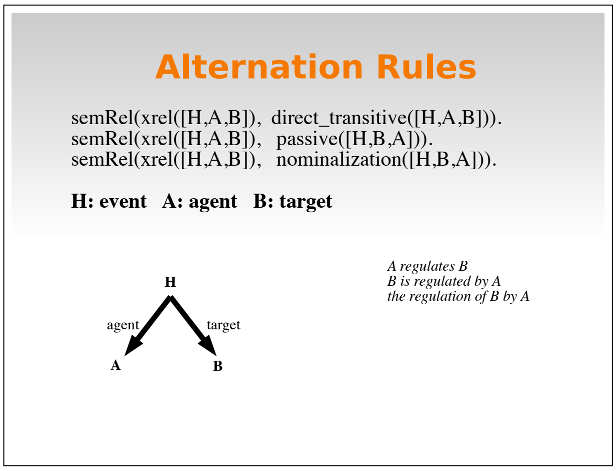


Syntactic Rules

```
synRel(passive, [X2,X1,X3],
 [ dep(subj,X2,X1),dep(pobj,X2,X3),
   dep(prep,X3,By), pos(X2,'VBN'),
   lemma(By,['by','through','via']) ]).
```

A regulates B
B is regulated by A
the regulation of B by A

Figure 2.2: Syntactic Rule for *passive verb-agent-target* structure



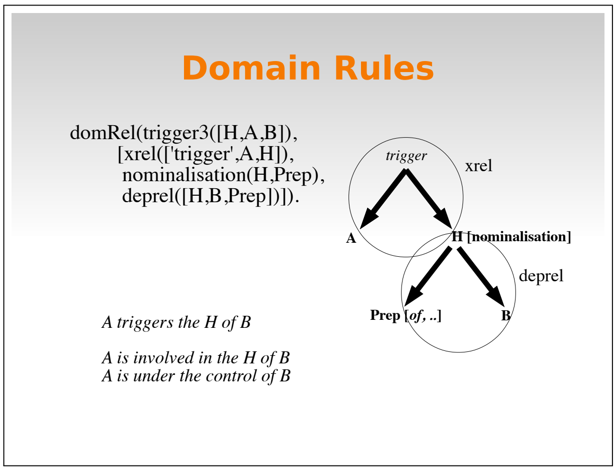
Alternation Rules

```
semRel(xrel([H,A,B]),  direct_transitive([H,A,B])).
semRel(xrel([H,A,B]),   passive([H,B,A])).
semRel(xrel([H,A,B]),   nominalization([H,B,A])).
```

H: event A: agent B: target

A regulates B
B is regulated by A
the regulation of B by A

Figure 2.3: Alternation Rule for transitive, passive, and nominalisation



Domain Rules

```
domRel(trigger3([H,A,B]),
      [xrel(['trigger',A,H]),
       nominalisation(H,Prep),
       deprel([H,B,Prep])]).
```

A triggers the H of B

A is involved in the H of B
A is under the control of B

Figure 2.4: Domain Rule for the the event verb *trigger*

hapax legomena: there is basically no bound to the indirectness in which a relation can be expressed, and indirect expression is rather the norm than the exception. The second column, $p(relevant)$ contains the probability of a path to be relevant, as described in section 2.3.1. Most paths have relatively low probabilities, which is partly due to the fact that only interactions which are novel and of key importance in an article are annotated in this challenge (which in turn is one of the reasons why we use additional features, see section 2.3.2). Except for the caveat that only novel interactions were annotated, and except for the lack of an explicit mapping of alternations onto each other, this data forms an alternation model. It does not directly express the probability of an alternation (its productivity), but it expresses the probability of the various configurations to mean the same specific thing, namely to express a relevant interaction between two arguments. Productivity can be calculated from the patterns, though: lines with a shared head lemma are typically the alternation variants.

I have summarised criticisms of traditional approaches to verb alternation in section 1.3.4. In particular, there are usually strong interactions between various factors instead of a binary decision by the language user. The detection of the envelope of variation is typically not data-driven, often not clear, typically not binary, and dependent on other decisions. Arppe et al. (2010) dismisses the classical concept of alternations as a whole.

Let us remember that the core definition of alternations is that they are meaning-preserving operations. E.g. Kreyer (2010, 169) defines them semantically, as variations exploiting the fact that “language can provide different syntactic means to express the same propositional content, to convey the same information” and first gives non-binary examples. A major problem with the view of alternations as binary, freely switchable choices is that there are complex interactions between alternations and that productivity is seriously reduced. At a first sight, a syntactic, precision-based definition seems to offer a more stable ground than a semantic definition. But low productivity entails that blindly applying alternations leads to a precision problem: applying the alternation very often leads to unacceptable utterances. We have seen in section 1.3.4 that for the case of the genitive alternation, 84% of the Saxon Genitives could also have been rendered as an of-PP without a strong change in meaning, but only 38% of the of-PPs could equally have been expressed with a Saxon Genitive. At the same time, the presence of complex interactions entails a recall problem: applying one alternation is typically not enough to find all semantically equivalent alternatives.

One possible solution is to take the semantic definition of meaning-preserving alternation as a base. All possible ways of expressing the same thing are the results of combinations of alternations. If we accept the assumption that all possible ways of expressing the same thing are due to alternations, than we obtain a recall-based definition of alternations. I will refer to them as *semantic alternations*. If we find a large, carefully annotated resource which marks the expression of the same underlying semantics, then we can build a data-driven model.

The well-resourced domain of biomedical relation mining can provide the data for such a model. The aim of IR generally is to find as much information that is semantically as closely related to a query as possible. In relation mining, the information to be found typically consists of the triple of the event and its two arguments. All events of the same type express the same core meaning, and training resources aim at being complete: all relevant interactions are annotated. This approach is presented in further detail in Schneider and Rinaldi (2011). While this suggestion is linguistically elegant, it also comes with practical problems. As we can see in Figure 2.1, the frequencies are much lower, data is much sparser, than in our investigation of the dative shift in section 1.3.2. Even in the well-resourced biomedical domain, there is not yet enough richly annotated data, but the situation is improving each year.

When applied to relation mining tasks, automatically learnt data-driven patterns as those shown in Table 2.1 typically lead to higher recall. This is the case because they include patterns and alternation interaction types that introspective approaches would not easily detect (as example 20). One such alternation type, which one could call the *sort-of* alternation, maps long forms involving a pre-qualifier (like *group of* or *sort of*) to short forms. When I annotated the GENIA data in Schneider et al. (2009) I observed that semantically lightweight nouns mostly do not influence the decision whether a path is relevant, the truth-functional semantics ‘shines through’. As a test if a relation is expressed by an example sentence, I often paraphrased it, using the top node and the two terms. The

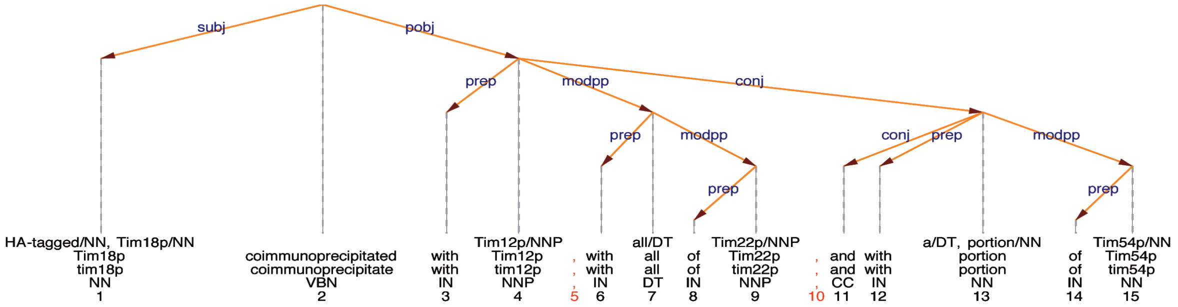

subj
pobj
prep
modpp
conj
prep
modpp
prep
conj
prep
modpp
prep
HA-tagged/NN, Tim18p/NN Tim18p tim18p NN 1
coimmunoprecipitated coimmunoprecipitate VBN 2
with with IN 3
Tim12p/NNP Tim12p tim12p NNP 4
, , , 5
with with IN 6
all/DT all all DT 7
of of IN 8
Tim22p/NNP Tim22p tim22p NNP 9
, , , 10
and and CC 11
with with IN 12
a/DT, portion/NN portion portion NN 13
of of IN 14
Tim54p/NN Tim54p tim54p NN 15

Figure 2.5: Dependency parser analysis of the sentence *HA-tagged Tim18p coimmunoprecipitated with Tim12p, with all of Tim22p, and with a portion of Tim54p.*

sentence *A activates groups of B* essentially expresses that *A activates B*, or *A blocks activation of B* expresses that *A blocks B*, whereas *A activates C, which has a binding site for B* does not express that *A activates B*. Words like *group* and *activation* are prototypes. I would like to use the term *transparent words*, which was originally coined by Meyers et al. (1998). Transparent words more often appear inside a path than outside a path. I thus simply calculate a score which divides its frequency inside a path by its total frequency. Words above a threshold are treated as transparent, and the subpaths which they are heading are accordingly shortened in the training and in the application phase. The two syntactic relations *conj* (conjunction) and *appos* (apposition) are equally semantically light-weight or even void, and therefore also cut.

We used these techniques in a back-off architecture to improve relation mining. The details of this approach are described in Schneider et al. (2009), we just give an example here. In the sentence in Figure 2.5 cutting conjunctions has the effect that *portion of Tim54* appears at the same level as *Tim12*, and cutting the transparent words *portion* and *all* has the effect that *Tim54* also appears at the same level as *Tim12*, and *Tim22* is only one PP-attachment (*modpp*) lower. These cutting steps are essentially text simplification procedures, the have the effect that the chance of including a decision for a path increases considerably.

Alternations, also in the broadest possible sense, i.e. including transparent words and appositions, and the concept of semantic alternation, are discourse-related features. I have discussed them as such in some of my research, e.g. Schneider et al. (2012). These and other discourse-level concepts can improve relation mining, as I discuss in the following section.

2.5 Linguistic Discourse

In the previous section, I have presented transparent words as syntactic simplifications preserving the truth function of the discourse. Let us now turn to discourse-level concepts in which the unity and the structure of the document are essential.

Discourse is defined as “a unit of language larger than a sentence and which is firmly rooted in a specific context ” (Martin and Ringham, 2000, 51). Discourse is a broad area of linguistics, partly overlaps with pragmatics and includes a wide range of aspects, for example anaphora resolution, text genre studies, cohesion, felicity, and community-wide background knowledge. An important unit of measurement in discourse is the document. When John Sinclair writes “trust the text” (Sinclair and Carter, 2004) he also means trust the text as an important unit for disambiguating the words that occur in it.

Scientific documents have a largely standardised structure. Title, abstract, introduction, background methods, results, conclusion and reference section can be found in most documents. Title, abstract and conclusion section are particularly important for the description of relevant interactions. Boosting the interaction scores from these section, and punishing those from the background section, led to consistently better results. Such an approach,

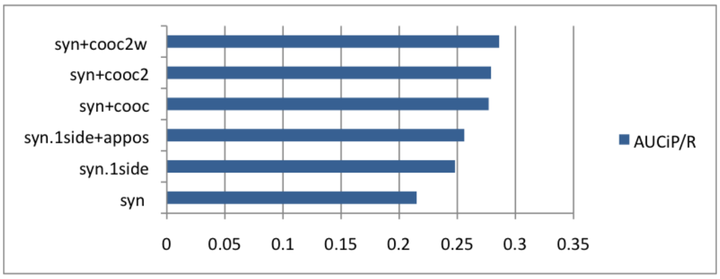


Figure 2.6: Performance on CTD relation ranking measured by AUCiP/R

typically called *zoning*, is standard for any IR approach.

Document-specific approaches to term recognition are used much less frequently, and we could show that they can increase the performance of relation mining. Ambiguity of terms is particularly high for acronyms. But it is rare that two acronyms in the same document have different meanings, as the one-sense-per-discourse hypothesis states (Gale, Church, and Yarowsky, 1992). One typically refers to the process of mapping an acronym to its long form as *expansion*. Different meanings of acronyms almost always coincide with two different expansions. We use two approaches to expanding acronyms depending on the document context, as follows.

2.5.1 Expanding Introduced Acronyms

Schwartz and Hearst (2003) introduce an approach for detecting acronyms in brackets. I use the syntactic relation *apposition* instead (which includes brackets, but contains more), and additionally profit from the reference which the term recogniser finds. My algorithm adds the expansion to all acronyms that are introduced in a document, if their references differ. This step increases recall at the cost of precision. I illustrate the algorithm with an example.

(21) The current studies were designed to examine if quinone intermediates are involved in the toxicity of hepatotoxic halobenzenes, bromobenzene (BB) and 1,2,4-trichlorobenzene (1,2,4-TCB). (CTD, pubmed 10092053)

The acronym *BB* in sentence 21 is erroneously tagged with a reference to the semantic type *gene* by our term recognizer, while it is an acronym of a *chemical substance*, namely *bromobenzene*. The fact that *BB* is connected via a syntactic *apposition* relation, in this case a bracket, makes it obvious for the reader that *BB* refers to *bromobenzene*. In scientific discourse, abbreviations are typically introduced in this fashion. The algorithm thus assigns the type chemical substance, referring to *bromobenzene* to all 5 occurrences of *BB* in the document.

The improvement in performance on CTD is summarised in Figure 2.6. Details are given in Schneider et al. (2012, Table 2). The evaluation metric AUCiP/R measures the area under the interpolated Precision/Recall curve, and assesses the quality of the ranking. The meanings of the labels are as follows. **syn** is purely our syntactic method, as described in section 2.3. **syn.1side** uses half-path features (from the event to term$_1$, and from event to term$_2$ independently, see section 2.3.2) as a backoff. The fact that this step improves performance shows us that data sparseness is serious. **syn.1side+appos** additionally recognizes acronyms that were introduced by a syntactic apposition relation, as just explained. The algorithm improves performance. **syn+cooc** works as follows: the sometimes low recall of syntactic methods can be increased by including sentence-coocurrence, which on its own is a frequently used baseline feature: whenever two entities appear in the same sentence, they are added as relevant interaction with a low score. This step increases recall at some cost to precision. We see, however, that simple surface methods often work quite well. Adding advanced syntactic methods often can only add the icing on the cake, improving a high baseline only very modestly, as we will also see in section 2.5.4. We also noticed that many interactions are expressed across several sentences. In **syn+cooc2**, the sentence-coocurrence score was thus extended to include the neighbouring sentence. Recall increases, but is compensated by an equally large decrease in precision. **syn+cooc2w** weighs the sentence-coocurrence score by distance, giving higher scores to entities that appear closer. Simple IR search operators such as $NEAR$ may appear to be linguistically irrelevant, as they have no syntactic motivation, but in fact they have a psycholinguistic explanation: close entities are more easily accessible.

2.5.2 Filtering Acronyms without Expansion Candidates

We also developed an approach in which those concepts of short acronyms that do not have promising expansion candidates in the document are filtered out. The algorithm checks the list of terms found in the document against the list of variants of terms in the reference terminology. We have used maximum entropy (linear regression) models with different features. The detailed results are given in Schneider et al. (2012, Table 3), but we can summarise as follows. First, dependency path features improve relation detection. Second, **Appos**, our expansion of acronyms that were introduced by a syntactic apposition relation, as on CTD, improves relation mining, while **filtering acronyms** introduced in this section leads to better precision and F-score, but recall suffers. Third, cutting **transparent** words leads to a slightly higher performance.

Filtering acronyms increases precision at the cost of recall. If all acronyms used in an article were introduced properly as style guides suggest, it would only lead to increased precision. However, very many acronyms are not introduced in the document, so that the increase in precision is smaller than the loss in recall. The lesson we can learn from this is that even texts in structured and regulated domains do not abide to set best practice standards. We will see the same effect in the following section.

2.5.3 Named Entity Recognition as Document Classification Task

We have seen in the two previous subsections that the context of the document can disambiguate the specific named entity class of acronyms thanks to the one-sense-per-discourse hypothesis (Gale, Church, and Yarowsky, 1992). Non-abbreviated named entities are sometimes also ambiguous (one string can map to several references), and very often there is considerable variability in them (one reference can be expressed by a large variety of strings). Particularly multi-word terms can e.g. show alternations, changes in word order, use of different synonyms or incomplete terms. We have participated in a research challenge aiming to detect the use of scientific methods in biomedical relation discovery. Details are given in Schneider, Clematide, and Rinaldi (2011). Here we show that the Firthian hypothesis (Firth, 1957), which we discuss in section 3.1.3, improves named entity recognition.

Scientists use a standardized set of assays and complex methods to prove interactions between proteins and genes. Extensive dictionaries with synonym lists exist, so that I first used a classical string-matching algorithm as is standardly done in named entity recognition. However, only 10-15% of the methods sections of all articles in this research challenge contained matches to dictionary entries, which means that such an approach will necessarily have very low recall. The vast majority of method terms are multi-word terms, for example MI:0004 has the term *affinity chromatography technology* and the dictionary synonyms *affinity chrom* and *affinity purification*. Subsets or recombinations of these strings are often used in research articles instead of the official dictionary term or a given synonym. I first use submatches at the word level. For instance, an occurrence of the word *purification* in the methods section of an article marks it as a candidate for MI:0004. About 80% of the methods sections of all articles contain word submatches, i.e. single words from the multi-word entries. Recall of a system using term submatches is higher, but precision is low. The problem is that, on the one hand, some submatch words are contained in many different experimental methods (they do not discriminate well), and on the other hand, that many submatch words very often do not refer to a method. For example, MI:0231 has the term name *mammalian protein protein interaction trap*, which means that every occurrence of the word *protein* assigns a score to this method.

To respond to these observations, a statistical method can be used. We use, on the one hand, conditional

Method termness		$p(method\|word)$		
Probability	term word	Probability	word	method
0.831	anti	0.490	L1	MI:0006
0.692	pooling	0.470	LT	MI:0019
0.662	hybrid	0.447	ERK1/2	MI:0006
0.519	x-ray	0.443	hydrogen-bonding	MI:0114
0.515	coimmunoprecipitation	0.441	omit	MI:0114
0.484	coip	0.438	synapses	MI:0006
0.469	bret	0.436	tumours	MI:0006
0.396	fret	0.435	REFMAC	MI:0114
0.369	tag	0.430	p21	MI:0006
0.367	tomography	0.424	COOT	MI:0114
0.356	bifc	0.423	epithelium	MI:0006
0.354	diffraction	0.418	flower	MI:0018
0.329	resonance	0.417	IKK	MI:0006
0.322	epr	0.412	caspase-3	MI:0006
0.322	crystallography	0.407	NF-kB	MI:0006
0.312	two-hybrid	0.406	floral	MI:0018
0.311	2-hybrid	0.406	9.00E+10	MI:0007
0.307	itc	0.404	diffracted	MI:0114
0.307	spr	0.403	atom	MI:0114
0.303	biosensor	0.403	HIV-1	MI:0007
0.300	two	0.401	wwwpdborg	MI:0114
0.300	saxs	0.401	CCP4	MI:0114
0.296	bimolecular	0.396	BK	MI:0006
0.296	plasmon	0.396	FRET	MI:0055
0.283	bait	0.394	MCF-7	MI:0006
0.282	fluorescence	0.394	contoured	MI:0114
0.282	nmr	0.390	Å	MI:0114
0.272	isothermal	0.389	hypoxia	MI:0006
0.258	calorimetry	0.387	c-Myc	MI:0007
0.258	one-hybrid	0.387	PI3K	MI:0006
0.247	crosslink	0.385	specification	MI:0018
0.238	tap	0.385	seed	MI:0018

Table 2.2: Termness and $p(method|word)$

probabilities for the method given a submatch word $p(method|submatchword)$ and, on the other hand, the probability of a submatch word to be a term, its "method termness". The method termness probability is measured as the conditional probability that a word occurence is actually part of one of the method terms of the document, i.e. where the annotator has assigned a method that contains the word to the document $p(termword = yes|submatchword, document)$. For example, 83% of the occurences of the word "anti" come from documents where methods containing "anti" (e.g. *"anti bait coip"*) have been used.

The performance of the system is evaluated in detail in Schneider, Clematide, and Rinaldi (2011), but one can summarise as follows: using $p(method|submatchword)$ and $p(termword = yes|submatchword, document)$ improves performance considerably. When discarding the dictionary completely, using not only submatch words but any word, thus essentially conducting a bag-of-words unsupervised document classification, the performance drops again to the level of the term submatch approach.

Best results are obtained when both methods are combined: partly using the dictionary submatches, and partly document classification. An excerpt of method termness probabilites $p(termword = yes|submatchword, document)$ is given in Table 2.2, left hand side. An excerpt of frequent words $p(method|word)$ including words that are not term submatches in the dictionary is given in Table 2.2, right hand side. Many of the words indicating experimental methods at high probability are cues to the expert for a particular method. For example the Ångström measure $\AA$ $(1\AA = 1^{-10} meters)$ is mainly used in some methods. We have learnt from this task that, for at least some named entity tasks, discourse in the form of the context given by the entire document is as important as a dictionary of the entities to be found. This is particularly so for those terms which can be expressed in many ways, those which consist of several words and those which are subject to considerable syntactic variation. The context of the document disambiguates sufficiently for the human reader, and an automatic classifier can partly compensate by using document classification.

2.5.4 Discourse Frequency as Salience

I have mentioned that we use various features in addition to syntactic features, for example the zoning features that I have mentioned in the previous section. A trivial feature measuring the salience of a term is simply counting

	Precision	Recall	F-Score	AUCiP/R
Standard	0.103	0.323	0.116	0.180
Feature 1 only	0.045	0.314	0.064	0.155
Feature 2 only	0.038	0.305	0.058	0.051
Feature 3 only	0.038	0.305	0.058	0.080
Feature 4 only	0.039	0.305	0.059	0.136
Feature 5 only	0.112	0.325	0.122	0.220
Surface only	0.145	0.319	0.148	0.191
Feature 5 only, surface	0.145	0.309	0.144	0.182
Feature 5 only, syntax	0.112	0.325	0.122	0.220
Reduced syntax backoff	0.119	0.329	0.124	0.196

Table 2.3: Performance of combined, single and revised features in our BioCreative ii.5 participation

its frequency in the unit of discourse, typically the document. The chance that a given document describes an interaction between the two most frequently mentioned terms is indeed very high. For our participation in Biocreative ii.5 (Rinaldi et al., 2010b) we used the following features:

1. Syntactic path: a version of our syntactic feature, as described in section 2.3.
2. Known interaction: Interactions that are already reported in the IntAct database receive a low score. The older the entry data in the database, the lower the score.
3. Novelty score: On the basis of linguistic cues (e.g., *"Here we report that ..."*), we attempt to distinguish between sentences that report the results detected by the authors from sentences that report background results. Interactions in "novelty" sentences are scored higher than interactions in "background" sentences.
4. Zoning: The abstract and the conclusions are the typical places for mentioning novel interactions, whereas the introduction and methods section are less likely and get lower scores.
5. Pair salience: Proteins that are mentioned frequently in an article are more likely to participate in a relevant interaction than proteins that are mentioned only once. We use the following simple calculation to assign a value to this feature: $sal(p_1, p_2) = \frac{f(p_1)*f(p_2)}{f(proteins\ in\ article)}$

We also measured the performance of each feature on its own for an error analysis. The results are given in Table 2.3. As expected, feature 1 performed better than features 2 to 4, but the fact that the simple frequency-based feature 5 performed extremely well, better than the standard system using all 5 features, surprised us. We need to bear in mind that feature 5 depends on the syntactic feature as well, in that only candidate pairs which are syntactically connected (in *any* way) are used. In order to remove this influence, we used a version in which only the surface strings (intended as backoff) generate pairs, see line *feature 5 only, surface* in Table 2.3. The low recall indicates that many patterns are missed now. We also noticed that some of our extensive syntactic backoff features had detrimental effects. After removing them we get the *reduced syntax backoff* system in the last line in Table 2.3.

We learn that simply counting the occurrences of the candidate proteins, as feature 5 does, seems to be one of the best methods for ranking candidate protein pairs. Due to the many repetitions of the core interactions, chances that at least one of them can be retrieved with a surface method are very high (this explains the good performance of the surface backoff). In other words, the simple discourse feature of frequency places a very high baseline, as frequency and salience are very closely correlated.

Another important conclusion that can be drawn from these results, which are also illustrated in Table 2.3 is that syntactic methods do provide a beneficial contribution. For a semi-automated annotation scenario, which is one of our goals, improved recall is often more useful than increased precision.

2.5.5 Semantic Expectations

Research on discourse often focusses on the expectation that the context (such as social context, or genre, or the document) places. I have not yet integrated a context-specific ontology to model expectations, but I have added a

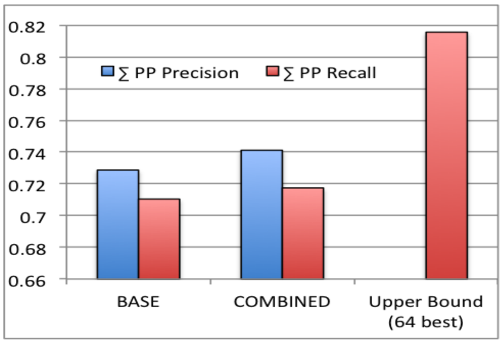


Figure 2.7: Results of evaluation on PP-attachment relations, with (COMBINED) and without (BASE) semantic expectations and distributional semantics

global semantic expectation model to the earlier, purely syntactic, expectations of the Pro3Gres parser. The details are described in Schneider (2012b).

The original parser models probabilities using only those syntactic relations that are in competition. For example, objects (e.g. *eat pizza*) and nominal adjuncts (e.g. *eat Friday*) are modelled as being in competition, because they have the same surface configuration. But subjects and objects are not seen in competition, as they hardly ever have the same surface configuration. The original parser strictly models syntactic competition. I have then added semantic competition: every relation is in competition with every other relation. In order to calculate the probability for a verb-object relation between *rabbit* and *chase* I also respect how many subject-verb relations between *rabbit* and *chase* the training corpus contains. This has the effect that a sentence like *the rabbit chased the dog* is assigned a lower probability than *the dog chased the rabbit* because rabbits are very unlikely to be subjects of active instances of *chase*. Thus, our semantic world knowledge becomes part of the model, the parser parses for what is semantically more plausible in cases of ambiguity. In computational terms, this has the same effect as extending the derivation model ($p(subject\ attachment|chase, rabbit)$) with a generation model ($p(subject(chase, rabbit)|chase)$. While such an approach entails the risk of misinterpreting surprising new information, it is also psycholinguistically adequate: human parsers often disambiguate by using their expectations and their world knowledge, and experience garden path situations if they are given totally unexpected information.

I have also added distributional semantics to alleviate sparse data in PP-sequences. I have used non-negative matrix factorisation (Lee and Seung, 2001) and obtained further improvements. The impact of the semantic expectation model and distributional semantics (COMBINED) on the PP-attachment relations, compared to the original parser (BASE) and the upper bound (if an oracle picked the best of the top 64 parses), is summarised graphically in Figure 2.7, more detailed tables are given in Schneider (2012b).

Schneider (2012b) introduces two further semantic modules: first, better informed decisions are given higher scores than less informed decisions, which improves performance. Second, I have added a self-training component, using the parsed 100 million British National Corpus (BNC) to alleviate sparseness. It marginally improves performance.

We can learn from my experiments that using semantic information improves parsing, and that in some cases data sparseness is worse than the errors introduced in the automatic method. The former is another indication that semantic expectations also help us to disambiguate, the latter that up to a point a parser can improve itself.

2.6 Intermediate Conclusions

I would like to summarize the linguistically relevant findings gleaned from the previous discussion as follows.

First, we have shown that using syntactic approaches improves Relation Mining, for example in Schneider, Kaljurand, and Rinaldi (2009) and Schneider et al. (2012).

Second, I have shown in Schneider (2012b) that semantic expectations help the syntactic parser, and that the automatically parsed data can be adduced to improve the performance of the parser (section 2.5.5). These are applications in which the parser, which was developed on linguistic rules, can give us linguistic insights and ultimately improve itself: the linguistic insight that semantic expectations help us to disambiguate has been used to improve the parser.

Third, I have shown that respecting transparent words improves Text Mining performance in Schneider, Kaljurand, and Rinaldi (2009), as seen in sections 2.5.1 and 2.5.2. In essence, my treatment of transparent words is a basic text simplification approach, which leads to simpler patterns and thus less sparse data. This text simplifying procedure could also be termed the *sort-of* alternation because the long and short variant are semantically largely equivalent, and *sort of* and *group of* are prototypical cases.

Fourth, we also use additional linguistic insights. We use full paths or half-paths where possible, respecting the idiom principle, which states that the sequence of words and subtrees is essential. We do not use part-of-speech tags as features, as they are largely redundant.

Fifth, I have discussed the importance of discourse and the unity of the document in section 2.5.

- We have seen that the simple discourse feature of frequency places a very high baseline (Rinaldi et al., 2010b), see section 2.5.4. This indicates that frequency and salience are very closely correlated. While for the detection of collocations the confirmation that frequency-derived measures and idiomaticity are closely related (section 1.2) was expected, this came as a total surprise to us: the discourse-level feature of frequency of a term or term-pair in a document performs almost as well as our entire Text Mining system. On the other hand, the fact that salience and frequency are so closely related explains why TFIDF works quite well to detect keywords and topics.
- I have also shown that a document-based approach to the expansion of acronyms improves Text Mining (Schneider et al., 2012) in section 2.5.1. Human readers can correctly expand highly ambiguous acronyms based on discourse expectations.
- We have seen in Schneider, Clematide, and Rinaldi (2011) which is summarised in section 2.5.3 that tasks which may seem to be lexical named entity recognition (NER) tasks, in our task the detection of scientific methods, profit from being partly addressed as document classification task, profiting from the Firthian hypothesis. As the document context contains important cues, and the variation of method terms is too strong, approaches profiting form the unity of the discourse of the document can significantly improve the task. In psycholinguistic terms, the cues are priming factors which influence reader expectations.
- We have experienced repeatedly that simple surface methods often work quite well. Adding advanced syntactic methods often can only add the icing on the cake, improving a high baseline only very modestly, for example in 2.5.1, where sentence-coocurrence and nearness were important features. Simple IR search operators such as $NEAR$ may appear to be linguistically irrelevant, as they seem to have no syntactic motivation. While nearness has no motivation in a pure syntax-principle view, in a pure idiom-principle view it would be a very accurate measure. In fact they have a psycholinguistic explanation: close entities are more easily accessible. We turn to psycholinguistic models in the next chapter.

Sixth, I have discussed in section 2.4 that alternations may profitably be studied from the semantic rather than from the syntactic end, based on Schneider and Rinaldi (2011). Patterns that are used to express the same relation in a Text Mining perspective should be mapped to each other. As alternations are highly interdependent and decisions non-binary and plagued by limited productivity, a Text Mining approach may be more promising than a syntactic approach. In a Text Mining approach, language models play a central role. This discovery is essentially a psycholinguistic insight, which is the topic of the next chapter. Due to the last insight, we need to discard my playful coinage of the *sort-of alternation* again or only use it as a working metaphor, because the status of alternations is contested. Instead I suggest to use language models centrally in the following chapter.

Chapter 3

A Psycholinguistic Model

Lexico-grammatical phenomena like collocation, lexical preferences, discourse-based reader expectations and argument structure (such as alternation choices), are concepts that can appropriately be addressed by psycholinguistic approaches, and which ultimately have a psycholinguistic explanation. We have seen in section 1.1 that psycholinguistics was one of the crucial sources for the move to lexicalist approaches to syntax. Tomasello (2000) has shown that the vast majority of chidrens' early language uses item-based linguistic schemas prior to developing abstract schemas. Pawley and Syder (1983, 193) argue that also grown-up language users exploit language creativity only to a very small degree, "that native speakers do not exercise the creative potential of syntactic rules to anything like their full extent, and that, indeed, if they did do so they would not be accepted as exhibiting nativelike control of the language."

There are indications that avoidance of creativity can be observed at all levels of language. While Tomasello (2000) and Pawley and Syder (1983) deal with the interaction of lexis and grammar, similar observations have been made in semantic domains, for example in the study of politeness and of compliments. Politeness research such as Watts (2003) point out the importance of formulaic language. Terkourafi (2001, 187) states that "in the data collected, formulaic speech carries the burden of polite discourse. This finding ... raises the possibility that the use of formulae may be a prominent feature of polite discourse in any culture". In compliment research, Manes and Wolfson (1981, 115) state that "one of the most striking features of compliments in American English is their almost total lack of originality". In their investigation, three surface syntactic patterns accounted for 85% of the 686 compliments in their American corpus, and nine patterns cover 97.2% of their data. The most frequent pattern is

(22) NP *is/looks* (*really*) ADJ

where NP is a noun phrase, *is/looks* can be any copular verb, (*really*) stands for a facultative intensifier, and ADJ for any positively charged adjective. An example for this pattern is

(23) Your coat looks really great!

We applied these nine linguistic patterns to the BNC and evaluated them in Jucker et al. (2007). Precision turned out to be very low in some patterns. Based on our evaluations we can estimate that the 100 million word BNC contains 1000 to 2000 compliments. The most frequent pattern dominates the distribution even more than in Manes and Wolfson (1981). These findings confirmed three things to us. First, that we see a Zipfian distribution like in lexis also in the use of compliments; out of the almost infinite range of possibilities to make a compliment, only very few patterns are used regularly, creativity is largely absent. Second, most uses of compliments seem to be based not on creativity, but on past experience, compliments heard or read in the past. Thus they can very appropriately be treated with the help of the discipline which learns from and then applies past experiences, statistics. Third, the choice on the side of a language generator (the speaker or writer) is in fact very restricted (only few, semi-fixed patterns, Zipfian distribution, idiom principle), while the low precision of some patterns indicates that the correct "choice" of semantically possible interpretations between ambiguous structures is difficult to make if we use syntactic patterns – lexical interactions are needed to disambiguate. This entails among others that the need for a statistical language model is more acute for language analysis (which is a $1 : many$ mapping problem) than for language generation (which is a $1 : few$ mapping problem).

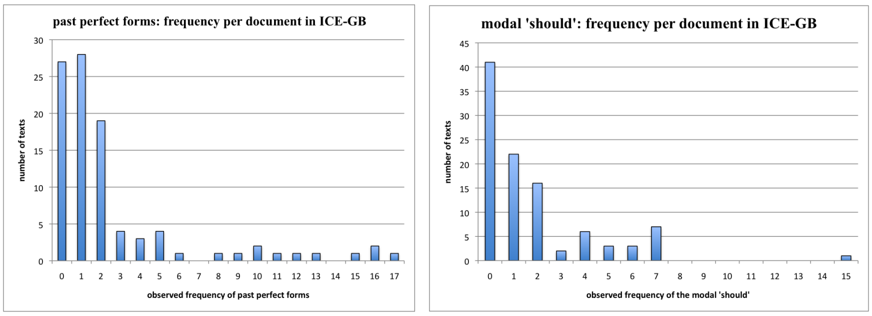


Figure 3.1: Per-document distribution of past perfect forms, and of the modal *should* in a large subset of ICE-GB

3.1 A Case for Statistical Language Models

Statistical models are very widely used in science, for example in meteorology, architecture, engineering, or physics. The use of multifactorial, predictive models has increasingly been recognised as important in Descriptive Linguistics, e.g. regression models (Evert, 2006; Gries, 2010; Gries, 2012). Descriptive linguistic approaches still standardly use significance testing instead of statistical models to describe linguistic differences. Let us first discuss in section 3.1.1 why that is not sufficient as only one factor is observed, before we turn to the important factor of genre or subgenre, and after this we will see that there is a strong correlation between word-word combinations found in corpus data and psycholinguistic expectations by readers or listeners.

3.1.1 Significance Testing for Descriptive Linguistics is not Enough

Significance tests, which are important to test if an observed difference between two varieties or genres or speaker groups is big enough to be statistically significant rather then just random variation, face the following serious problems, as some of the assumptions which they make are often not met. These are the assumptions.

1. Assumption of random distribution
2. Assumption of independence from other factors
3. Assumption of free choice

Assumption of random distribution The T-test makes the assumption that data is distributed normally. If this were the case, the distance from one occurrence of a given word to its next occurrence would be normally distributed, and so would be its frequency per document. However, Church (2000) shows that the probability for a word to occur is much higher if it has already been seen in the same document than if it has not appeared before, and hence jeopardizes the assumption of random distribution. This is another consequence of the importance of discourse: a document tends to have a core topic, and a coherent structure, choice of lexis, and a limited set of participants. The one-sense-per-discourse hypothesis is also a consequence of this. Simple IR measures such as TFIDF use this fact, and I have just discussed how closely frequency and salience are related. Evert (2006) shows that non-random distribution of words does not only apply to content words, but also grammatical features, passive forms in his case. It has been claimed that the use of the passive form is lower in American than in British English (e.g. Leech et al. (2009)), but the number of passive forms per document is not distributed normally. This non-randomness can be observed for a variety of linguistic entities. We have shown in Schneider and Hundt (2012) that the distribution of past perfect forms and of modal verbs is equally non-random. Figure 3.1 shows that their distribution per document in a large subset of ICE-GB is very different from a Gaussian bell-shaped distribution.

Sedlatschek (2009) claims that past perfect form is significantly more frequent in Indian English, using a χ^2-Test on ICE India. He cautions that significance is low, and that it mainly applies to the earlier Kolhapur corpus.

1. Modal Verb ← Variety $\star$ Genre :
   Genre*** $>$ Variety$\cdot$
2. Modal Verb ← Variety $\star$ Genre + Text :
   Text*** $>$ Variety** $>$ Genre

Figure 3.2: Significance ordering of factors of the modal verb regression pilot study

Whatever the outcome of the test, the non-randomness of the distribution has to lead to unreliable significance test results. Nelson (2003) claims that the modal *should* is significantly more frequent in East African English, using a χ^2-Test, on ICE data. I got the same result, but again, due to the questionable presumptions that are made in applying the tests, the validity of this conclusion is questionable.

Assumption of independence from other factors While the parametric T-test assumes normal distribution, the non-parametric χ^2-Test does not make such an assumption. The assumption taken in χ^2 is that the data is independent from other factors. In Figure 3.1, the document itself is a strong factor (which again highlights the importance of document-level discourse, see section 2.5): the chance of drawing a past perfect form is significantly higher in some documents than in others. In this sense, an independent token is only generated by each individual text; the number or the percentage of the forms under investigation found in that document is its scalar measure. But also the individual texts are largely influenced by various factors: on the one hand classical sociolinguistic factors like age, gender, social background of the author, and on the other hand the topic, and the genre of the document, leading to different reader expectations and background knowledge, which is for example essential for word-sense disambiguation.

Gries (2010) also discusses an example where the application of monofactorial analysis, and multifactorial analysis without considering all interactions, can lead to incorrect results. He concludes that "multifactorial data must be analyzed multifactorially: ... the complexities of linguistic data do not reveal themselves easily either to the naked or to the monofactorial eye" (Gries, 2010, 143).

Assumption of free choice The third problem is that a speaker often has no real choice. The assumption of free choice as I have named it here is strictly speaking not a statistical assumption, but a consequence of the influence of factors like topic and semantics of the document, and in turn often leads to the problem of non-randomness of distributions. The absence of choice is most obvious for content words, where the lexical range is largely predetermined by the semantic topic. But also when it comes to choosing function words and other morphosyntactic factors, for example tense, voice, aspect and modality, discourse and content often place restrictions. In a narrative, the simple past will be more frequent, in a scientific paper, the resultative present perfect is often used, in the weather forecast we expect to hear future tense. In a discussion on moral ethics, modal verbs are inherently very frequent. At best, the speaker in such a discussion has a choice between several modal verbs of obligation (*must, should, need*). One should approximate the speaker choice as envelope of variation (Labov, 1969; Sankoff, 1988), see section 1.3.2, as closely as possible.

3.1.2 Pilot Studies

Bresnan et al. (2007) use a logistic regression model to predict the dative alternation using a rich set of features. They predict up to 95% of the binary decisions, their model only has a 5% residual. I have conducted a similar GLM model pilot study on the written parts of ICE-GB, NZ, India, Fiji, and Ghana, for the choice of modal verb among *must, should, need* to test the influence of the factors *genre* and *individual text*. The findings, using asterisk notation to indicate significance levels of the factors in the model, are summarised in Figure 3.2[1]. I list the

[1] Significance levels are standardly indicated as follows: $p = 0$ as '***' ; $p < 0.001$ as '**' ; $p < 0.01$ as '*' ; $p < 0.05$ as '·'

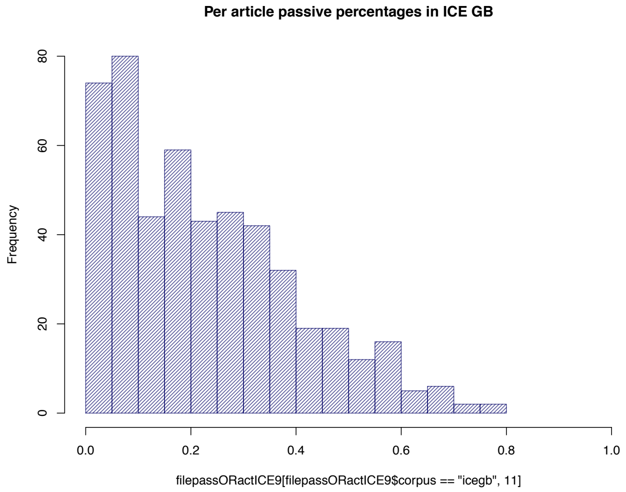


Figure 3.3: Per article passive percentages in ICE GB

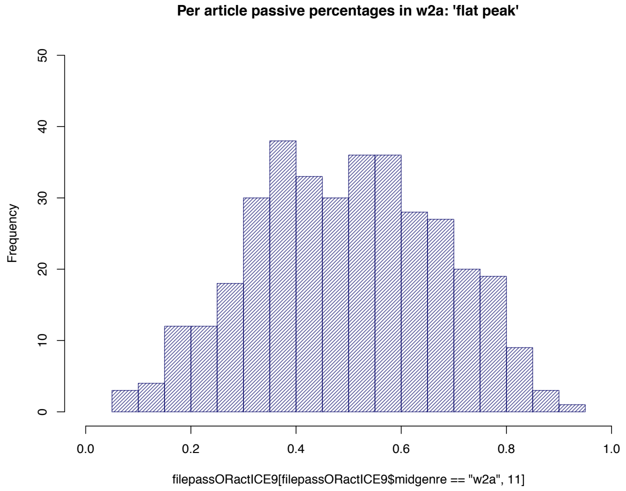


Figure 3.4: Per article passive percentages in ICE GB scientific

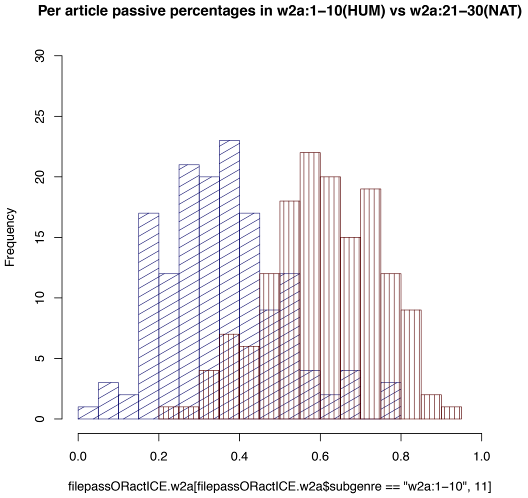


Figure 3.5: Per article passive percentages in ICE GB scientific subgenres HUM and NAT

factors in order of significance. For each occurrence of a modal verb of obligation (*must, should, need*) the model attempts to predict which of the three should be used, given the three simple factors *Variety* (ICE-GB, NZ, India, Fiji, Ghana), *Genre* (scientific writing, press, fiction, ...) and, in the second model, also the document ID *Text*. In the first model, Genre is more significant than Variety. This indicates that Genre is a stronger factor than Variety, as e.g. Evert (2006) describes for passive voice, which is particularly frequent in the scientific genre, and which leads to the effect that small Genre differences overshadow possible Variety effects. Interactions are also reported as highly significant.

In the second model, the factor of the individual Text is even more important than the Genre. This indicates that the topic may be an even more important factor than Variety. The result that Genre is not significant according to the model does not mean anything, because there is a complete dependence from Text to Genre: if the document ID is known, its genre can be predicted deterministically. The fact that the individual text is such an influential factor may indicate, *in extremis*, that unless we have parallel corpora we cannot compare any texts, as their topic will be different. In more moderate terms, it indicates that one needs a level below the genre, such as subgenres or topics. Which is the right level for modelling the reader expectations between genre, which is too coarse, and document topic, which is too detailed and ridden by sparse data?

Inspired by Evert (2006) we have investigated the passive data in the Brown family and in the ICE corpora, using syntactically parsed data (Schneider, 2008; Lehmann and Schneider, 2012a). We compare passive forms to transitive verb forms, and exclude verbs which exclusively exist in only the active or the passive form, in order to approximate the envelope of variation of the passive alternation. Figure 3.3 shows passive percentages in ICE-GB, per article. The leftmost bar, for example, indicates that 77 documents contain fewer than 5% passive forms. The randomness assumption is clearly violated, the distribution is far from a Gaussian normal distribution. Fiction texts typically contain very few passives, scientific articles relatively many. In Figure 3.4 we compare texts from the scientific genre. We get a peak around 50%, the distribution looks a bit more similar to a normal distribution, but it is too flat around the peak – in fact it is a bimodal distribution. When we split the scientific genres (ICE W2A) into their four given subgenres *Natural Sciences (NAT)*, *Social Sciences (SOC)*, *Humanities (HUM)*, and *Technology (TEC)* each subgenre shows a distribution which is approximately normal. Figure 3.5 shows the distribution in the subgenres *Humanities (HUM)* (which contains least passives) and *Natural Sciences (NAT)* (which contains most passives). We now have a reasonable normal distribution, in all subgenres, hence we do not miss important factors and can confidently apply significance testing. The differences between the genres, according to the T-Test, are highly significant.

The level of subgenre seems to be the right level between genre and topic, in this case. Descriptive Linguistics percentage expectations, and article reader expectations, are probably best expressed at this level for the passive alternation. If one does not have detailed subgenre annotation, or if it turns out that we need a different level, Evert (2006) suggests to use a data-driven method for the detection of subgenres or very broad topics, using distributional semantics based on vector space models.

3.1.3 Frequencies and their Psycholinguistic Reality

Vector space models are increasingly used in Computational Linguistics. They are data-driven approaches learning topics, concepts, and synonyms from purely distributional aspects of large text collections. The distributional hypothesis, which is typically attributed to Firth (1957) (it is sometimes also called the Firthian hypothesis) and to Harris (Harris, 1968; Harris, 1970) says that "words with similar distributional properties have similar meanings" (Sahlgren, 2006, 21), or in other words, that words which frequently co-occur are similar in meaning, that words which occur frequently in a document are closely related to its content, that words generally are largely defined by their context. Using frequency-derived measures of words and co-occurrences is neither a recent invention, nor a Computational Linguistics one. Central ideas of the distributional hypothesis can be traced back to Saussure's *différance*.

> The differential view on meaning that Harris assumes in his distributional methodology does not originate in his theories. Rather, it is a consequence of his theoretical ancestry. Although Harris' primary source of inspiration was American structuralist Leonard Bloomfield, the origin of the differential view on meaning goes back even further, to Swiss linguist Ferdinand de Saussure. Saussure (1857-1913) was one of the fathers of modern linguistics, and an important inspiration for the Bloomfieldian structuralism that developed into the distributional paradigm. (Sahlgren, 2006, 57)

Saussure saw the (arbitrary) connection between *signifiant* and *signifié* as a social convention. He was not mainly interested in how the convention arises, but relegated this question to diachronic linguistics, and focussed on the synchronic linguistic system, in which each word is defined (*valeur*) by being different from other functionally similar words. Such a differential definition of language is also used in vector-based models.

Data-driven approaches, including collocation detection, approximating the envelope of variation for alternations, describing genre and regional differences, finding relations in Text Mining, or applying vector space models, depend on corpus frequencies. But how far can one trust the correlation between frequency and psycholinguistic salience? If we want to learn anything about semantics or the structure of language from word distributions then it is essential that there is a strong correlation between word distributions and the linguistic intuitions of native speakers. In fact, there often are strong correlations between frequencies or frequency-based measures, and semantic concepts, often more strongly so then we anticipated. For example we have seen in section 2.5.4 that the brute-force feature 5 of simply counting pair frequencies performed almost as well as our entire Text Mining system for the detection of interacting entities. Further I have discussed in section 1.2 that although the substitution-based method intuitively seems like a better proxy to the semantic concept of non-compositionality than the similarity-based method, Wulff (2008) finds that the similarity-based approach performs better.

From a practical viewpoint, distributional measures are often the best simply obtainable and adaptable descriptions that one has. From such a practical viewpoint, also the question whether language arises from use or if the frequent co-occurrence is a by-product of the Saussurian social convention, can also be left underspecified, except for diachronic considerations.

Word-word combinations that are very frequent become entrenched and grammaticized and are stored as one formulaic unit in the mental lexicon. Schmid (2000) shows that there is a strong correlation between entrenchment and frequency. A typical effect of entrenchment is contraction. Krug (2003, 25) shows for example that frequency and contraction of pronoun-verb combinations are strongly correlated. Frequency of co-occurrence is not only an effect of entrenchment, it is also often described as a contributor, as functional linguists increasingly point out:

> Frequency is not just a result of grammaticization, it is also a primary contributor to the process, an active force in instigating the changes that occur in grammaticization ... I will argue for a new definition of grammaticization, one which recognizes the crucial role of repetition in grammaticization and characterizes it as the process by which a frequently-used sequence of words or morphemes becomes automated as a single processing unit. (Bybee, 2007, 337)

Reading time is also affected by formulaic language. Bod (2001) showed in a lexical decision task that high-frequency three-word sentences such as *I like it* are processed faster than low-frequency sentences such as *I keep it* by native speakers. MacDonald and Seidenberg (2006) gives an overview. Conklin and Schmitt (2012), a meta-study, confirms this, and also that non-native speakers find it more difficult to acquire formulaic language (as (Pawley and Syder, 1983) predicted), as they are exposed to fewer instances:

> Virtually every study, using a variety of research methodologies, shows that formulaic language holds a processing advantage over nonformulaic language for native speakers. However, for nonnatives, this is often not the case, although higher proficiency levels increase the chances of also enjoying this advantage. The crucial role of frequency in processing clearly applies not only to individual words

> but also to formulaic sequences. It appears that frequency of exposure is a key aspect of learning formulaic sequences. (Conklin and Schmitt, 2012, 56)

Further, eye-movement research shows that the fixation time on each word in reading is a function of the frequency of that word (frequent words have shorter fixations) and of the forward transitional probability (the conditional probability of a word given the previous word $P(w_k|w_{k-1})$, a bigram model). An information-theoretic version of such transition probabilities is known as *surprisal* (Levy and Jaeger, 2007). Variants of these are used by many computational linguistic tools, for example part-of-speech taggers. Psycholinguistically, eye movement experiments have shown that surprisal correlates to reading times (Demberg and Keller, 2008). We will come back to surprisal in section 3.2.4.

Gries and Wulff (2005) and Gries and Wulff (2009) find strong correlations between collocation strengths and experimentally obtained sentence completions from advanced L2 learners of English. Ellis and Ferreira-Junior (2009) find that frequency of learner uptake is predicted by frequency of occurrence, and even more so by collocation measures. Szmrecsanyi (2006) finds strong correlations between verbs' collostruction strengths and priming effects observed in different corpora and for different constructions. Frequency and entrenchment are very closely correlated, and frequency-based measures (such as collocation measures and surprisal), and salience and psycholinguistic expectations as well. Although frequencies and frequency-based measures are probably approximations to psycholinguistics, they are often better approximations than any other measures that we have, as e.g. Wulff (2008) shows concerning collocations.

So far, we have seen that on the *syntagmatic* level, i.e. concerning word sequences, corpus data and psycholinguistic data are closely correlated. In order to derive semantic knowledge, also the *paradigmatic level* needs to be considered, to establish if there is a correlation between common contexts and semantic similarity, as predicted by native speakers. Co-occurrence of words and human associations are often strongly correlated, as Schulte im Walde and Melinger (2008) discuss. They test how often psycholinguistic stimulus-response pairs are found in observation frames in large corpora. Synonyms and hypernyms can also be detected distributionally by constraining to syntactic relation instead of context window, e.g. (Curran, 2004). We have used distributional approaches on syntactic relation for the detection of biomedical taxonomy in Weeds et al. (2007). Rothenhäusler and Schütze (2009) show that approaches constraining to syntactic relation have considerably higher performance on the task of semantic classification of nouns. Using our large resources such as the parsed BNC will enable us to create genre-adapted semantic distributional semantic models in future research. The insight that such models are not only implementations of the Firthian hypothesis, but also consistent usage-based language models has not been fully recognised in linguistics. As a first step towards this direction, the recognition is slowly gaining ground that language is largely usage-based (Bybee, 2007), and that corpus-linguistics, cognitive science and psycholinguistics are tightly connected (Gries, 2012).

Gries (2012, 47) states that "cognitive approaches to language are not only compatible with much recent work in Corpus Linguistics, but also provide a framework into which corpus-linguistic results can be integrated elegantly." He suggests the use of examplar-based multidimensional models for words and their distributions He stresses that the distributional aspects and contexts are important factors. Vector-based models are examplar-based multidimensional models that may be ideal for this task: they can cope well with extremely sparse word-based dimensions, and they also allow us to include distributional and contextual information, most typically by including summed contexts (Schütze, 1998) in a Firthian sense. He concludes that research in the area is urgently needed.

> ... my main focus is the proposal for us corpus linguists to assume as the main theoretical framework within which to explain and embed our analyses a psycholinguistically informed, (cognitively-inspired) exemplar/usage-based linguistics. Thankfully, I am not alone in this. There are some linguists who have assumed at least somewhat similar positions already ((Schönefeld, 1999; Schmid, 2000; Mukherjee, 2004; Butler, 2004), for instance), but the major breakthrough I think is needed in order for corpus linguistics to shed its 'purely descriptive' label has not yet happened. The from my

> point of view most important arguments in a very similar spirit are from Miller and Charles (1991) as well as Hoey (2005). (Gries, 2012, 56-57)

In my research on psycholinguistic influences on language variation I aim to help redressing this shortage.

3.2 The Parser as a Psycholinguistic Model

Simplistic language generation models include monogram models, in which the choice by a user is to draw a word token from an urn of word types, the lexicon ($p(token)$), or n-gram models, in which a word ($token_0$) is drawn conditioned on the $n-1$ previous word tokens, e.g. for n=3:

$$p(token_0|token_{-1}, token_{-2}) \tag{3.1}$$

Part-of-speech taggers use n-gram models, in addition to token probabilities they also use tag probabilities. Such models can be seen as radical *idiom principle* models. The fact that they provide reasonable results on part-of-speech tagging tasks provides another hint to the importance of formulaic, sequence-based language use. But these models are insufficient as they fail to take any hierarchical structural information (which one usually calls syntax) into account. A list of the most frequent errors which taggers typically make also illustrates this shortcoming: in English, they include the distinction between prepositions and phrasal particles (Sag et al., 2001), in German the distinction between articles and relative pronouns (Volk and Schneider, 1998). While in a syntax-principle view, a speaker has to make choices at every node in the syntactic tree, in all non-radical *idiom principle* views, a user needs to take few syntactic choices for large subtrees. One such choice is the one between the different configurations in alternations, in the sense of envelope of variation (Labov, 1969; Sankoff, 1988), see section 1.3.2. In order to recognise the variant forms of an alternation, a parsing approach is essential.

3.2.1 Local and Global Models

I have mentioned in section 3.1.2 that Bresnan et al. (2007) use a logistic regression model to predict the speaker choice in the dative alternation using a rich set of factors (indeed involving animacy, pronominality, lexis, etc.) which allows them to predict up to 95% of the binary decisions correctly. It deals with one specific speaker choice and could thus be termed a *local model*. However, local models are partly insufficient, for the following reasons.

First, if an outcome depends on other outcomes, local models are insufficient. This is for example the case in persistence (Szmrecsanyi, 2006) or in the Saxon Genitive, where nested *'s* constructions are extremely rare. Also alternations are highly interdependent, as Arppe et al. (2010) point out and which has prompted my suggestion of semantic alternations in section 2.4. Syntactic relations are a prime sample of highly interdependent factors: if a noun is subject, it cannot be an object, if a PP is attached to a verb it cannot be attached to a noun, and correct disambiguation depends on other disambiguations. Local models are obviously of limited use, as they miss the context of other parsing decisions. A parser is a global syntactic model. A parser gives us the right tool for example to punish a nested Saxon Genitive, to give high or low scores to analyses that are supported by, or unlikely due to lexical preferences.

Wasow and Arnold (2003) have conducted a psycholinguistic experiment in which users were presented the following sentences:

(24) a. The foundation gave Grant's letters to Lincoln to a museum in Philadelphia.
b. The foundation gave a museum in Philadelphia Grant's letters to Lincoln.

(25) a. The foundation gave Grant's letters about Lincoln to a museum in Philadelphia.
b. The foundation gave a museum in Philadelphia Grant's letters about Lincoln.

Sentence 24 is ambiguous, the PP headed by *to* could be expected to attach to the verb first, leading to a garden path situation. Sentence 25 does not show this ambiguity. Automatic approaches to this experiment which do not use a parsing approach are doomed to fail in principle, as they are intrinsically unable to "see" and treat the ambiguity. Only a parser allows us to consider all factors, on the one hand the statistical features and on the other hand the parsing context.

Second, the fact that Bresnan et al. (2007) predict up to 95% of the binary decisions correctly shows not only that they have an excellent model, but also that the speaker in fact has relatively little choice, that the interconnection between the rich features involved may have a largely deterministic effect, that these features in combination may prime us almost completely, that the choices have largely been made in choosing the features, that even using a carefully constructed envelope of variation may fall too short. In parallel to lexis, which "is complexly and systematically structured and that grammar is an outcome of this lexical structure" (Hoey, 2005, 1), the features in interaction systematically predict the local outcome of the alternation. Alternations are then, equally, only an epiphenomenon.

I think there are two possible explanations for this. Either, we continue down this road into the direction of Sacks' *order at all points* (Sacks, 1995, 484), which "understands order not to be present only at aggregate levels and therefore subject to an overall differential distribution, but to be present in detail on a case by case, environment by environment basis." (Sacks, 1995, xlvi). Order at all points entails that all outcomes are deterministic once one has enough factors, and is an argument against the use of statistics. If you have all relevant factors, you can re-construct all others. This is what loss-free compression algorithms do. But as the approach requires all factors, and contradiction-free classification, the complexity of data is just shifted to an equally long set of factors. No insights are gained, and Occam's razor has been neglected. We need models performing non-loss-free (lossy) compression, up to a point.

The other explanation is to say that the speaker choices are not local but rather distributed across the selection of features and their interactions, and the interaction between the syntax and the idiom principle. In order to model psycholinguistic decisions, we need a *global model* which takes these interactions into account, combines and weighs local models, and uses the parsing context. This is precisely what a parser does.

I have mentioned at the beginning of this chapter that the need for a statistical language model is more acute for language analysis, which is a $1 : many$ mapping problem, than for language generation, which is a $1 : few$ mapping problem. For a parsing model, the language analysis perspective is the obvious choice: utterances that have already been generated are waiting to be analysed. Language analysis probability $p(analysis|words)$ and generation probability $p(words|analysis)$ are related, if we use Maximum-Likelihood Estimation the relation is given by Bayes' rule:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)} \tag{3.2}$$

Approaches using global models on psycholinguistic tasks are still rare. Outstanding exceptions are Borensztajn, Zuidema, and Bod (2008) who have measured L1 language acquisition complexity using a Data-Oriented Parser (DOP) model (Bod, 1992; Bod, Scha, and Sima'an, 2003) or Demberg, Keller, and Koller (2014), who use a similar Tree-Adjoining Grammar (TAG) approach. They use parser scores and lexical and structural surprisal to calculate global and local psycholinguistic processing difficulty. There is an enormous potential for using such approaches to predict ambiguity and variation and measure syntactic surprisal, and feed back the obtained knowledge into the parser. As Language variation aims to explain reasons why sentences are rendered in specific ways, there is an enormous potential for using parsers as psycholinguistic language models.

3.2.2 Back to the Idiom and Syntax Principle

I have discussed the idiom and syntax principle (Sinclair, 1991; Francis, 1993; Hunston and Francis, 2000) in section 1.1. Parsing can also be seen as a tug-of-war between the syntax principle, in which a *competence* grammar licenses possible analyses, and the idiom principle, in which *performance* preferences rank them and constantly prune (filter) unlikely structures. Computational Linguistics has used statistical models for a long time. When Bresnan et al. (2007) present their regression model they state that "We have found that linguistic data are more probabilistic than has been widely recognized in theoretical linguistics" (Bresnan et al., 2007, 28). In Computational Linguistics, the gradient and probabilistic nature of language has been a constant 'problem' as the disambiguation between a plethora of correct, but overgenerating rules has always been a major task for language analysis, for example in syntax, e.g. (Collins, 1999; Schneider, 2008; Nivre, 2006).

Schneider (2008) combines a hand-written *competence* grammar and probabilistic *performance* disambiguation learnt on the Penn Treebank. The blind application of syntax rules typically leads to between dozens and hundreds of analyses for real-world sentences. For disambiguation, bi-lexical preferences (Collins, 1999) can be used. Bi-lexical preferences, which can for example be implemented as bi-lexically conditioned probabilities, are the Computational Linguistics implementation of Hoey (2005, 1)'s insight that "lexis is complexly and systematically structured and that grammar is an outcome of this lexical structure". My parser (Schneider, 2008) uses Maximum Likelihood Estimation (MLE) to estimate the probability of the dependency relation R at distance (in chunks) $dist$, given the lexical head a of the governor and the lexical head b of the dependent. I have introduced the equation already in section 1.1.

$$p(R, dist|a, b) = p(R|a, b) \cdot p(dist|R, a, b) \cong p(R|a, b) \cdot p(dist|R) = \frac{f(R, a, b)}{f((\sum R), a, b)} \cdot \frac{f(R, dist)}{fR} \tag{3.3}$$

Parsers that are based on regression models (Maximum Entropy) also exist, e.g. (Charniak, 2000; Nivre, 2006). Lexcalized statistical parsers do exactly what Sinclair (1991), Francis (1993), and Hunston and Francis (2000) predicted. They allow us to measure the strength of the two principles in interaction (or what psycholinguists refer to as *routinization* and *chunking*). And other linguistically relevant factors can be added without limitations. I have added various new statistical, semantic, and discourse-level resources. For example, in Schneider (2012b) I have added semantic expectations (section 2.5.5). The underlying assumption for the parser to serve as psycholinguistic language model is proof by application: if the use of a feature improves parsing performance, then we can assume it to be (psycho-)linguistically influential.

The advantage of models is, in general terms, that they can take several factors into account and make predictions. The advantage of using a parser as a model is that it can supplement local models with a global model, combine a multitude of local models in their interaction, and make parsing disambiguation predictions. The following are reasons why we can treat a parser as a model:

- it takes attachment decisions (predictions) based on grammar rules and lexical preferences
- its statistical model can be extended by whatever factors that we observe
- it learns form real-word data: the syntactically annotated Penn treebank
- it aims to represent reality: deliver correct syntactic analyses
- the model fit is an important psycholinguistic measure: Entrenched structures get higher scores, as they are expected

In chapter 1, we have measured the signal which the parser emits and described the findings arising from this approach, gradient phenomena such as collocations, alternations and regional variation, symbolised by *Signal of real-world variation* in Figure 3.6. I have discussed in Schneider and Hundt (2009) that the parser signal is reasonably accurate, and that errors are dominated by white noise. Since the parser offers a language model to the researcher, one can also apply it in many different ways in addition to measuring the signal. One can modify model parameters, and use specific textual sources as input.

Minimally the model parameters indicated in Figure 3.6 can be changed in the following ways to describe the roles of ambiguity and disambiguation, formulaic language, language learner failures, readability, alternations, argument structure and lexical priming: on the parser's input side, one can change the competence *grammar*, add more *local models* (e.g. Schneider (2012b)), vary the input by issuing *L2 real world failure* data or forcing *random alternations* (voice, dative shift, synonyms, Saxon Genitive etc.), or *vary the statistical model*.

Let us next discuss experiments which are based on changing some of the parameters lined out in Figure 3.6.

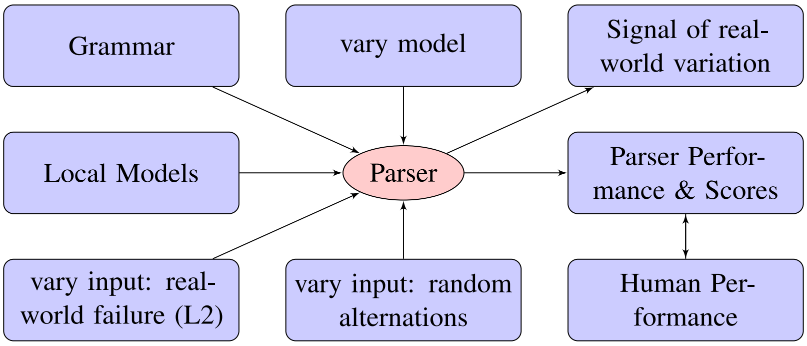
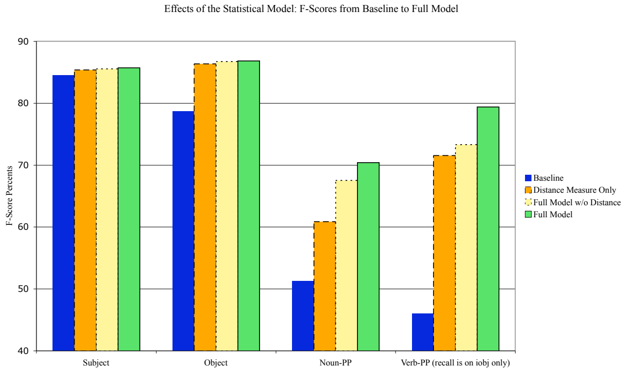

Grammar
vary model
Signal of real-world variation
Local Models
Parser
Parser Performance & Scores
vary input: real-world failure (L2)
vary input: random alternations
Human Performance

Figure 3.6: Selected input and output parameters of our language model



- Baseline: only syntax rules: w/o idiom principle
- Distance measure: recency
- Full M w/o distance: bi-lexical preferences
- Full M : recency + bi-lexical preferences

Figure 3.7: Gain in parser correctness due to probabilistic disambiguation by distance and bi-lexical preference

- One can vary the model: I show an example of applying the parser with and without lexical priming in Figure 3.7. The increase in parsing accuracy from the baseline offers a clear indication of the importance of lexical associations on human parsing and automatic parsing alike: on the one hand the increase on automatic parsing performance is considerable, especially on highly ambiguous relations, such as PP-attachment. On the other hand, the hypothesis that human parsing largely relies on lexical associations is supported.
- One can use real-world data with real-world failures, e.g. learner English (L2), as I will in the following section 3.2.3.
- Our use of a syntactic parser as global model on large corpora allows us to compare human and machine parses: do they make similar errors? Which sentences are difficult to process? Do the utterances of learners fit the model less well? I will address these questions in sections 3.2.3.
- One can randomly manipulate input using permitted syntactic operations, for example in order to test if alternations are possible, and if we avoid ambiguity (see section 3.2.5).

3.2.3 Varying the Input

As an example of varying the input, we use real-world learner data from a learner corpus containing original L2 and error-corrected utterances in Schneider and Grigonyte (2013). We use the error-corrected Japanese Learner English Corpus NICT (`http://alaginrc.nict.go.jp/nict_jle/index_E.html`). The corpus consists of

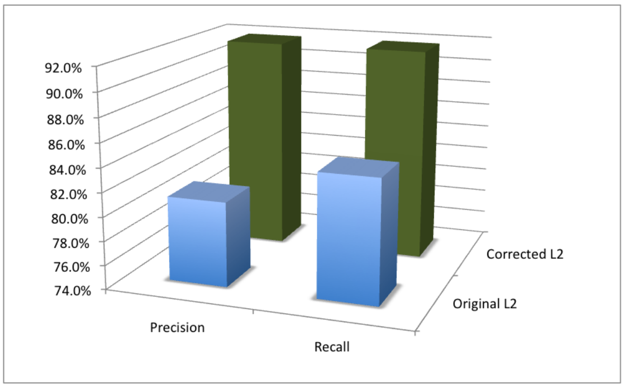


Figure 3.8: Performance of the parser on > 500 syntactic relations and their corrected counterpart

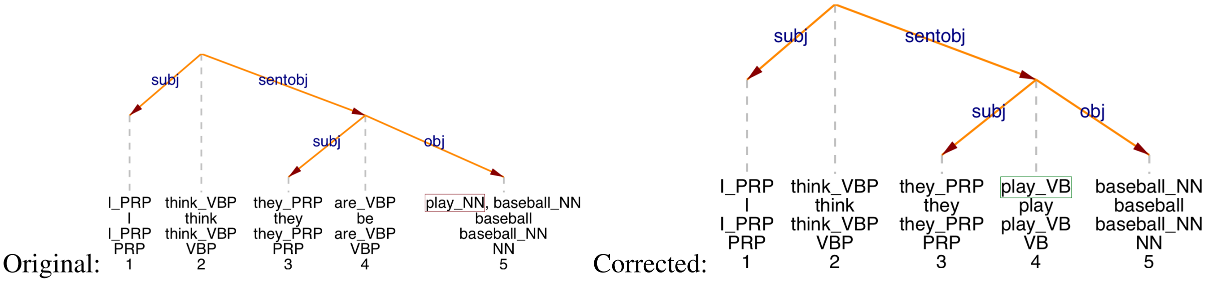

Original:
subj
sentobj
subj
obj
I_PRP I I_PRP PRP 1
think_VBP think think_VBP VBP 2
they_PRP they they_PRP PRP 3
are_VBP be are_VBP VBP 4
play_NN, baseball_NN baseball baseball_NN NN 5
Corrected:
subj
sentobj
subj
obj
I_PRP I I_PRP PRP 1
think_VBP think think_VBP VBP 2
they_PRP they they_PRP PRP 3
play_VB play play_VB VB 4
baseball_NN baseball baseball_NN NN 5

Figure 3.9: Parse of *I think they are play baseball* and its corrected counterpart

spoken data from exam interviews. Our hypothesis was that L2 utterances do not fit the model very well, as their language contains errors and formulaic sequences are used less in native-like fashion, as Pawley and Syder (1983) predicted and Conklin and Schmitt (2012) confirm. L2 utterances may thus give rise (for the human reader) to increased processing times and ambiguity, and (for the automatic parser) to lower performance and lower parser scores on the parser's output side. Millar (2011) confirms that L2 utterances and unexpected use of learner collocations do increase human processing times and ambiguity. This is a key psycholinguistic concern. He investigates the processing by native speakers of learner collocations which deviate from target-language norms. Results show that such deviations are indeed associated with an increased and sustained processing burden.

Parser Performance

Let us now investigate how the parser deals with L2 input, if it reacts in a way similar to what Millar (2011) reports. We have manually annotated > 500 syntactic relations from random sentences from the NICT corpus. Figure 3.8 shows that the error rate almost halved on corrected text. As expected, the parser performs less well on the original learner utterances, which contain many errors, but better on the same learner utterances after many of the mistakes have been corrected. An example is given in Figure 3.9. The mistagging of *play* as a noun in Figure 3.9 by the automatic parser does not mean that human parser would also make such an error. But the fact that the tagger, which is trained on large amounts of real-world context data, suggests a noun tag here indicates that the verb reading is surprising in this context, and human readers or listeners may show slightly increased processing load or minimal delays in comprehension.

V	Sentence	Score
ORIG	Usually , I go to the library , and I rent these books .	5054.31
CORR	Usually , I go to the library , and I borrow these books .	8956.83
ORIG	For example , at summer , I can enjoy the sea and breeze .	7186.86
CORR	For example , in summer , I can enjoy the sea and breeze .	8965.99
ORIG	so I will go to the Shibuya three o ' clock , nannda , before Hachikomae .	176.172
CORR	so I will go to Shibuya at three o ' clock , nannda , in front of Hachikomae .	12787.4
ORIG	The computer game is very violence in today , but I do n't like it .	6570.44
CORR	Computer games are very violent today , but I do n't like them .	161.753

Table 3.1: Parser scores of original and corrected sentences from the NICT corpus

Length	CORR	ORIG	CORR / ORIG
1 – 5	134	116	1.15
5 – 10	20272	17513	1.16
10 – 15	2121509	1504041	1.41
15 – 20	1563909630	895645699	1.74
20 – 25	16999039476	11879067407	1.43
25 – 30	42255013311	30951544002	1.37

Table 3.2: Parser scores of original and corrected sentences in NICT, and ratios, by sentence length in chunks

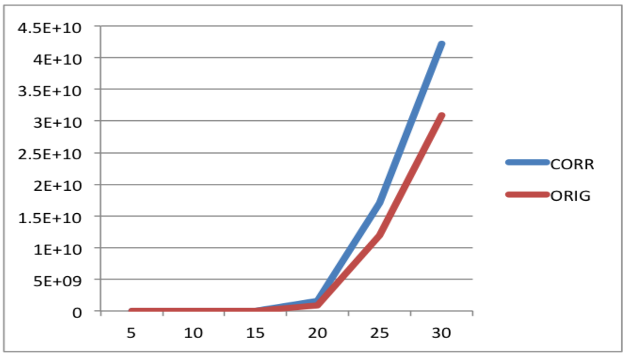


Figure 3.10: Scores of original and corrected sentences in NICT, by sentence length in chunks

Parser Scores

Probability-based scores which are originally intended for disambiguation and ranking of parsing candidates can be used as measures of surprise and model fit. A high parser score indicates:

- the utterance matches the expectation, and a particular syntactic parse
- the lexical items as seen in combination strongly point to a certain analysis

A low parser score indicates:

- the utterance is unexpected by the model
- the parser cannot map it well to any syntactic analysis

Keller (2003) has shown that there can be a strong correlation between grammaticality and parser scores. In German, "SOV is generally regarded as the basic word order for subordinate clauses. Verb initial orders are regarded as ungrammatical." Keller (2003) compared human parser magnitude estimation scores against automatic parser probability, and showed that they strongly correlate.

The examples in Table 3.1 compare parser scores of original and corrected sentences from the NICT corpus. In the last sentence, the corrected version obtains a lower score, but this is partly due to the fact that scores depend on sentence length. The means and the ratio (score of corrected / score of original) are given in Table 3.2. A

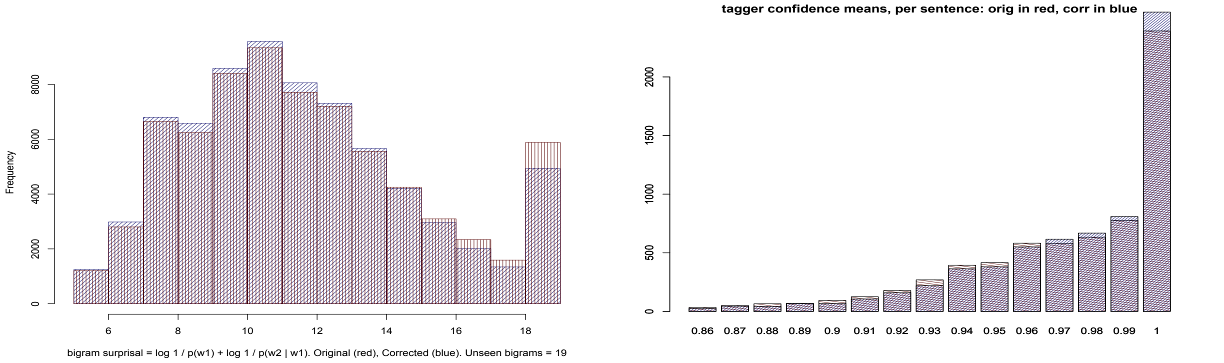


Figure 3.11: Distribution of bigram surprisal (left) and tagger confidence (right), compared between original L2 data (in red) and corrected L2 data (in blue) of the NICT corpus

graphical comparison of scores by sentence length (measured in chunks) is given in Figure 3.10. We can see that the corrected learner utterances obtain considerably higher scores, which confirms our hypothesis that they fit the parser model better, which correlates to the findings in Millar (2011), indicating that the parser processes language in a similar way to humans in this respect.

Many of the mistakes that are corrected in the NICT corpus are not just subtle misuses of collocations, which characterise the difference between a reasonable and a native-like command of English, in the sense of Pawley and Syder (1983), but grammatical mistakes. Furthermore, also some of the error corrections made in the corpus are questionable, and not all errors are corrected.

In order to have data from a less controlled setting, we also wanted to compare learner English from different learner levels. Our hypothesis is that utterances of speakers and writers from more advanced levels obtain higher parser scores, as they fit the model better. To test this, we have compared parser scores of the written CEEAUS (Corpus of English Essays Written by Asian University Students, (Ishikawa, 2009)). This corpus is not error-corrected, but the subject matter is tightly controlled, and fine-grained learner levels are provided in the metadata. The learner competence level relations (lower $<$ middle $<$ semi-upper & upper) are as we had hypothesized.

3.2.4 Surface Models: Surprisal and Tagging

Model fit of L2 data tends to be lower at all levels. While using a parser measures model fit at the level of interaction between idiom and syntax principle, also pure idiom principle models which only consider the surface sequence of words and word classes show reduced model fit, even if much less strongly than the parser model. The distribution of surprisal (Levy and Jaeger, 2007) for corpus bigrams, which is calculated as $log\frac{1}{p(w_{n-1})} + log\frac{1}{p(w_n|w_{n-1})}$ has a mean of 11.5 (and standard deviation of 3.36) for the NICT corrected text, and 11.7 (and standard deviation of 3.48) for original learner text. The distribution is given in Figure 3.11 on the left. I have obtained the surprisal probabilities from the BNC.

The higher mean indicates higher surprise about the continuation of the utterances, the higher standard deviation indicates that uniform information density (Levy and Jaeger, 2007) has been observed less. Uniform information density arises from the tug-of-war between expressivity (speakers want to convey information) and formulaicity (speakers observe the idiom principle) and leads to a normal distribution. Levy and Jaeger (2007) propose that it can be seen as a principle minimizing comprehension difficulty.

Also part-of-speech taggers rely on a surface language model which uses transition probabilities. Each tagger decision has a certain probability. In the language model of the tree-tagger (Schmid, 1994), the mean of the confidence probability of the tagger for the original L2 data is 0.9686, for the corrected data it increases to 0.9712.

In other words, also for the tagger, surprise is higher, model fit is lower. Averaged per sentence we get the frequency distribution shown in Figure 3.11 on the right. An important difference between the surface model of surprisal and the parser model is that while the former measures the idiom-principle surprise of the continuation, the latter measures the entropy and hence difficulty of the attachment decisions taken during parsing, or in other words the ambiguity. The fact that this ambiguity model shows differences between the original and the corrected L2 data more clearly (compare Table 3.2 and its graph in Figure 3.10 to Figure 3.11) bears particular promise.

3.2.5 Ambiguity

A prototypical case of ambiguity are garden path sentences. In them, a discrepancy between a *local maximum* and a *global maximum* exists: a locally most plausible interpretation needs to be revised due to subsequent text data. Garden path situations are exemplified by sentences like *"the horse raced past the barn fell"*. Garden path situations mean that a locally most plausible interpretation needs to be revised due to subsequent text data so that a globally possible or at least more plausible interpretation can be found. In a statistical parser this is conveyed by a locally relatively unlikely interpretation becoming the most likely at a later stage.

I have shown in Schneider et al. (2005) that an automatic parser deals with this ambiguity similarly to the human recipient. An experiment using short beam lengths leads to a situation in which locally least plausible analyses get lost. Without a repair mechanism such as backtracking the globally correct interpretation cannot be reached when using the short beam. I used 6 alternatives per span as a normal beam and 2 alternatives per span as a short beam. In our 500 sentence evaluation corpus (Carroll, Minnen, and Briscoe, 1999), although true garden path sentences are rare, 13 sentences get less correct analyses in the short beam scenario, for example

(26) Mitchell$_1$ said$_2$ [the Meiner administration]$_3$ and$_4$ [the Republican]$_5$ controlled$_6$ [State Senate]$_7$ share$_8$ [the blame]$_9$

Comparing the parse chart entries reveals that an object relation between *controlled* (at position 6) and *Senate* (at position 7) is about 20 times more likely than an adjective relation. The chart spans from *Republican* (position 5) to *Senate* (position 7) leading to the correct global span additionally include the rare nchunk relation – a relation that corrects chunking shortcomings. The chart entry containing a subject relation to Republican and an object reading to Senate is 210 times more likely than the nchunk plus adjective entry that leads to the globally correct span. If aggressive pruning such as short beam is used at this stage, no global span can be found by the parser: the parse fails, corresponding to a situation that triggers a human parser to re-analyse.

The fact that sentences which could lead to garden path situations are rare, can be interpreted as an indication that we try to avoid such situations.

(27) He saw the flower I like.

(28) He saw the flower pots like.

While sentence 27 is perfectly acceptable, also native speakers may find 28 difficult to understand. First, *flower pots* is a strong collocation which triggers its interpretation as a single chunk, one noun phrase. Second, on semantic grounds, it is very unlikely that non-animates like pots are subjects of the verb *like*. Third, pronouns as in sentence 27 cannot be pre-modified, which means that the presence of a zero-relativizer is likely.

The question whether one avoids ambiguity is still contested. Wasow and Arnold (2003) reject the suggestion that argument ordering plays a major role. However, they only investigated one alternation, in a lab setting. Humans easily disambiguate, but they may need a little bit longer or show increased processing load.

Wasow and Arnold (2003) also admit that they hardly found the ambiguous structures that they were looking for, also in large corpora. The fact that garden path situations are rare could be due to two reasons: either orderings which have the potential to lead to ambiguity are rare, or one indeed avoids them. In order to test this, I have added a component to my parser which randomly applies syntactic alternation operations and idiom replacements, leading to different argument ordering. If native speakers find the reordered sentences more difficult to interpret and the

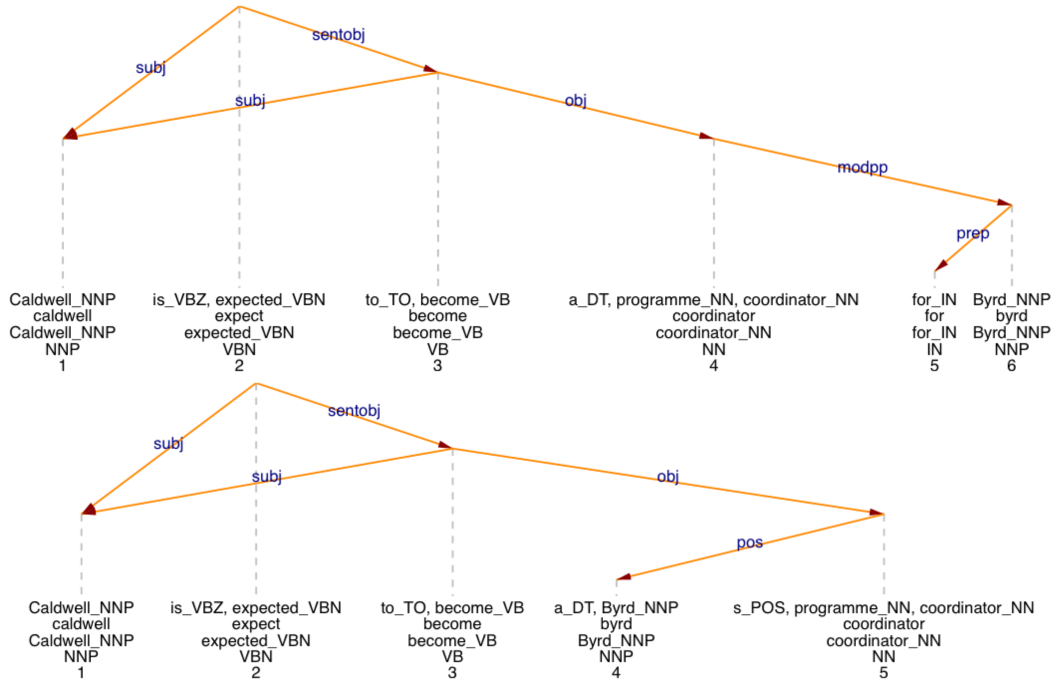

subj
sentobj
subj
obj
modpp
prep
Caldwell_NNP caldwell Caldwell_NNP NNP 1
is_VBZ, expected_VBN expect expected_VBN VBN 2
to_TO, become_VB become become_VB VB 3
a_DT, programme_NN, coordinator_NN coordinator coordinator_NN NN 4
for_IN for for_IN IN 5
Byrd_NNP byrd Byrd_NNP NNP 6
subj
sentobj
subj
obj
pos
Caldwell_NNP caldwell Caldwell_NNP NNP 1
is_VBZ, expected_VBN expect expected_VBN VBN 2
to_TO, become_VB become become_VB VB 3
a_DT, Byrd_NNP byrd Byrd_NNP NNP 4
s_POS, programme_NN, coordinator_NN coordinator coordinator_NN NN 5

Figure 3.12: Sample sentence before and after application of a random alternation, here the Saxon genitive alternation

automatic parser shows higher error rates, then the chosen ordering is purposeful. We compare psycholinguistic measures to parser scores and the increase of automatic parser errors. Figure 3.12 shows an example of a sentence with a random alternation. I have not measured parser scores or conducted large-scale analyses with this approach, but I plan to do so in future research.

3.3 Intermediate Conclusions and Outlook

I have argued for the use of statistical global models for Descriptive Linguistics. The advantage of models, generally, is that they allow us to take a multitude of factors and their interactions into consideration. Syntactic relations are a prime sample of highly interdependent factors, correct disambiguation depends on other disambiguations. Local models are of limited use, as they miss the context of other outcomes, i.e. parsing decisions. I have explained the difference between local and global models in section 3.2.1. Syntactic relations are a prime sample of highly interdependent outcomes: if a noun is subject, it cannot be an object, if a PP is attached to a verb it cannot be attached to a noun. Like human language users, a parser is a global syntactic model and gives us the right tool for example to punish a nested Saxon Genitive, to give high or low scores to analyses that are supported by, or unlikely due to lexical preferences, and to combine a multitude of local models in their interaction, and make parsing disambiguation predictions respecting the syntax principle (using an explicit grammar) and the idiom principle (using lexical interaction statistics).

I have argued that a syntactic parser can be used as a psycholinguistic global language model. I have shown that there is a correlation between the automatic parser and human parsers. In section 3.2.3 I have summarised key findings of Schneider and Grigonyte (2013):

- parser performance is considerably lower on the original utterances of language learners than on the corrected utterances

- also parser scores on corrected utterances are higher, and there is s a correlation between learner level and parser scores
- higher parser scores indicate better model fit, which indicates that advanced learners produce utterances that are more expected by the parser as language model

While model fit of the uncorrected L2 data is also lower for surface language models, the parser model shows more pronounced differences between the parser scores for the original and the corrected texts. The parser model is a model of ambiguity rather than surprise. In section 3.2.5 I have discussed a key finding of Schneider et al. (2005): an automatic parser deals with the ambiguity in garden-path sentences in similar ways to the human recipient.

I am interested on the one hand in how a parser can be used as a model of human language processing (section 3.2). On the other hand I am also interested in how much the parser can be improved by adding insights gained by learning how humans process data (e.g. section 2.5). While this imposes a certain danger of a cyclical argument, most tool and model development efforts are a dialectic incremental process, spiralling up to increased refinement levels based on the interaction between data and theory.

I plan to extend this approach in future applications. I will add further features to the parser, for example entropy and surprisal (Levy and Jaeger, 2007). Keller (2004) shows that entropy and processing effort are correlated, using a corpus of eye-tracking data. Such entropy-based measures can also serve as a base for the detection of errors by language learners (Gamon, 2011; Leacock et al., 2014), which I would like to integrate into the parser.

Even though it can be argued that “all models are wrong, but some are useful” (Box and Draper, 1987, 424), that statistical models fall short, they perform an important task in data compression to help us in interpreting linguistic data, which is often very complex. Although compression can typically not be loss-free, models deliver the most influential factors to us, allowing us both to predict new instances and interpret the data. Even though it can be argued that a parser is a relatively poor approximation to human performance, syntactic parsers offer language models which integrate a large number of features, respect the tug-of-war between the idiom and the syntax principle, and can be extended almost without limitation to include more and more appropriate features like surprisal, or semantic and discourse features. These possibilities will be explored in future research. Psycholinguistically adequate parsing models are sought for. Keller (2010) writes: “The challenge facing researchers in computational and psycholinguistics therefore includes the development of language processing models that combine syntactic processing with semantic and discourse processing. So far, this challenge is largely unmet”.

Chapter 4

Conclusions

This synopsis has addressed the research question of how Descriptive Linguistics, Text Mining, and Psycholinguistics can profit from the use of computational linguistic methods, and how the linguistic insights thus gained can be adduced to improve the Computational Linguistics tools again. The question gave rise to four guiding questions, which I very briefly take up again for a summary in the following paragraphs. More detailed conclusions are given at the end of each chapter, I only give an extremely brief summary here.

The question whether automatically parsed data, despite a certain level of errors, can deliver detailed results and insights that are useful for the study of linguistic variation and lexico-grammatical preferences was mainly addressed in **chapter 1**. We have seen that large amounts of parsed data deliver detailed insights into lexical interaction. I have described our experiments with using up to 1000 million words in sections 1.2 and 1.3, delivering both frequent and rare collocations and alternations. We have also seen that the list of collocations is open-ended. We have seen in section 1.2 that even in the frequent subject-verb and verb-object constructions, strong lexical and collocational preferences are at play, as Hoey (2005) predicted. In section 1.3, we have seen that both highly productive alternations like passive and restricted alternations like dative shift or Saxon genitive are marked by strong preferences throughout.

Concerning the noise produced by parsing errors, which may affect the validity of results, I have shown in section 1.4.2 that the signal of syntactic variation is stronger than the systematic skew which errors may introduce: parsing errors are largely white noise. Automatic parsing can also be applied profitably to diachronic linguistics for some tasks. Error rates are considerably higher on earlier texts, but I managed to reduce them by adding further resources and semantic expectations, as described in section 1.5.2. The same semantic knowledge also improves parsing of PDE texts, as I explained in section 2.5.5. This result has both a computational linguistic and a psycholinguistic interpretation.

The question if one can detect new patterns in language by using data-driven approaches was answered in subsections of chapter 1. We have applied data-driven approaches for the detection of new verb-preposition structures in section 1.4.3, using tagged and parsed data, and in section 1.4.4 for TAM profiles, using chunked data. Although a considerable amount of human intervention and interpretation is required, we have detected regional variation patterns. I have shown that data-driven approaches can be used to approximate the envelope of variation (Labov, 1969; Sankoff, 1988), i.e. the actual subconscious choices that speakers make, in section 1.3.

The question of whether Descriptive Linguistics can profit from insights gained in Text Mining has been addressed in **chapter 2**. We have detected relevant patterns in the texts partly in a data-driven fashion, learning from lean document-level annotation. The patterns are very sparse. I have described how linguistic insights and data-driven approaches like transparent words can simplify the patterns in sections 2.3 and 2.4. At the same time, we have learnt in section 2.5 that pure frequencies, discourse features and semantic expectations are essential. They have allowed us to disambiguate acronyms in the document context, to improve named entity recognition by document classification, and to improve the parser due to semantic expectations. From a Text Mining viewpoint, our experiments allowed us to improve Text Mining; from a psycholinguistic viewpoint, they indicate that discourse-based reader expectations and semantic expectations play a crucial role. I have further argued that speaker choices in alternations, although I have managed to better approximate them with data-driven approaches, are not binary. In order to include more relevant factors, it is essential to complement research by language models. I argue in section 2.4 that data-driven language models used in Text Mining can give us a more realistic approach to alternations. In future research, I will apply the findings of this chapter to register analysis (e.g. (Biber and Conrad, 2009)) and media content analysis (Wueest, Schneider, and Amsler, 2014).

The question whether statistical language models can overcome more of currently frequent shortcomings in descriptive linguistic investigations, and what type of statistical models are most promising, is addressed in **chapter 3**. We encountered actual significance testing problems in section 1.4.4, and discussed additional problems and presented solutions in section 3.1. We have also seen that variation inside a genre is stronger than typically assumed, that more fine-grained levels are needed, be it the subgenre level as in section 3.1.2 or the document-level as in section 2.5. A major reason for using models (such as regression models) in linguistics is the multifactorial nature of the speaker choices, and also fact that the interaction between different factors is enormous. I argue that syntactic parsers can add essential information on the interdependence between decisions that local statistical models cannot do in section 3.2. Such global models integrating more and more statistical factors and often submodels do more than predict local choices of speakers, as they take the entire array of interactions of both the factors and the outcomes at production and disambiguation at decoding into consideration. While model fit of uncorrected learner English data is also lower for surface language models, the parser model shows more pronounced differences between the parser scores for the original and the corrected texts. The parser model is a model of ambiguity. I suggest to address the question up to which point parsers can be used as models of human reading, and give preliminary answers in section 3.2. I have shown that there is a correlation between psycholinguistic factors and parser-derived measures. I argue that parsers therefore have the potential to be used as psycholinguistic language models. Such parser-based models still have a number of restrictions, in particular the number of features that they respect is too small, more semantic and discourse features need to be added, training data is small, and the set of relevant factors is probably infinite. But models will improve. On a practical level, parser-based human language processing models may in the future help the development of automatic error-detection for language learners, while on a theoretical level they may lead to better approximations to human language processing.

References

Abney, Steven. 1996. Partial parsing via finite-state cascades. In John Carroll, editor, *Proc. of the Workshop on Robust Parsing at the 8th Summer School on Logic, Language and Information*, number 435 in CSRP, pages 8–15. University of Sussex, Brighton.

Altenberg, Bengt. 1998. On the phraseology of spoken english: The evidence of recurrent word combinations. In A. P. Cowie, editor, *Phraseology: Theory, analysis, and applications*. Oxford University Press, Oxford.

Arppe, Antti, Gaetanelle Gilquin, Dylan Glynn, Martin Hilpert, and Arne Zeschel. 2010. Cognitive corpus linguistics: five points of debate on current theory and methodology. *Corpora*, 5(1):1–27.

Aston, Guy and Lou Burnard. 1998. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.

Baron, Alistair and Paul Rayson. 2008. Vard 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham. Aston University.

Biber, Douglas. 2003. Compressed noun-phrase structures in newspaper discourse: the competing demands of popularization vs. economy. In *New Media Language*. Routledge, London.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press, Cambridge.

Biber, Douglas, Edward Finegan, and Dwight Atkinson. 1994. Archer and its challenges: Compiling and exploring a representative corpus of historical english registers. In Udo Fries, Peter Schneider, and Gunnel Tottie, editors, *Creating and using English language corpora, Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993*. Rodopi, Amsterdam, pages 1–13.

Bick, Eckhard. 2003. A CG & PSG hybrid approach to automatic corpus annotation. In Kiril Simow and Petya Osenova, editors, *Proceedings of SProLaC2003*, pages 1–12, Lancaster.

Bick, Eckhard. 2010. FrAG, a hybrid constraint grammar parser for French. In *Proceedings of LREC 2010*, Valletta, Malta.

Bod, Rens. 1992. A computational model of language performance: Data oriented parsing. In *Proceedings COLING'92*, Nantes, France.

Bod, Rens. 2001. Storage vs. sentence memory: Computation of frequent sentences. In *Proceedings CUNY-2001 Abstracts*, Philadelphia, Pennsylvania.

Bod, Rens, Remko Scha, and Khalil Sima'an, editors. 2003. *Data-Oriented Parsing*. Center for the Study of Language and Information, Studies in Computational Linguistics (CSLI-SCL). Chicago University Press.

Borensztajn, Gideon, Willem Zuidema, and Rens Bod. 2008. Children's grammars grow more abstract with age - evidence from an automatic procedure for identifying the productive units of language. In *Proceedings of CogSci 2008*.

Box, George E. P. and Norman R. Draper. 1987. *Empirical Model Building and Response Surfaces*. John Wiley & Sons, New York, NY.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*. Royal Netherlands Academy of Science, Amsterdam, pages 69–94.

Bresnan, Joan and Tatiana Nikitina. 2009. The gradience of the dative alternation. In Linda Uyechi and Lian Hee Wee, editors, *Reality Exploration and Discovery: Pattern Interaction in Language and Life*. CSLI Publications, Stanford, pages 161–184.

Burger, John D. and Sam Bayer. 2005. MITRE's Qanda at TREC-14. In E. M. Voorhees and Lori P. Buckland, editors, *The Fourteenth Text REtrieval Conference (TREC 2005) Notebook*.

Butler, Christopher S. 2004. Corpus studies and functional linguistic theories. *Functions of Language*, 11:147–86.

Buyko, Ekaterina, Elena Beisswanger, and Udo Hahn. 2012. Extraction of pharmacogenetic and pharmacogenomic relations – a case study using pharmgkb. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pages 376–387, Hawaii.

Bybee, Joan. 2007. *Frequency of Use and the Organization of Language*. Oxford University Press, Oxford.

Carroll, John, Guido Minnen, and Edward Briscoe. 2003. Parser evaluation: using a grammatical relation annotation scheme. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*. Kluwer, Dordrecht, pages 299–316.

Carroll, John, Guido Minnen, and Ted Briscoe. 1999. Corpus annotation for parser evaluation. In *Proceedings of the EACL-99 Post-Conference Workshop on Linguistically Interpreted Corpora*, Bergen, Norway.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the North American Chapter of the ACL*, pages 132–139.

Choueka, Yaacov. 1988. Looking for needles in a haystack. In *Proceedings of RIAO '88*, pages 609–623.

Church, Kenneth. 2000. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In *In Proceedings of the 17th conference on Computational linguistics*, pages 180–186.

Clegg, Andrew B and Adrian J Shepherd. 2007. Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics*, 8:24.

Cohen, K. Bretonnel and Lawrence Hunter. 2008. Getting started in text mining. *PLoS Computational Biology*, 4(4):e20.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Collins, Michael. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29:589 – 637.

Collins, Michael and James Brooks. 1995. Prepositional attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, MA.

Conklin, Kathy and Norbert Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics*, 32:45–61.

Curran, James R. 2004. *From Distributional to Semantic Similarity*. Doctoral thesis, Institute for Communicating and Collaborative Systems, University of Edinburgh.

De Cock, Sylvie. 2000. Repetitive phrasal chunkiness and advanced efl speech and writing. In Christian Mair and Marianne Hundt, editors, *Corpus Linguistics and Linguistic Theory*. Rodopi, Amsterdam.

Demberg, Vera and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Demberg, Vera, Frank Keller, and Alexander Koller. 2014. Parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 40(1).

Ellis, Nick C. and Fernando Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, 7:187–220.

Erkan, G., A. Ozgur, and D. R. Radev. 2007. Extracting interacting protein pairs and evidence sentences by using dependency parsing and machine learning techniques. In *Proceedings of BioCreAtIvE 2*.

Evert, Stefan. 2006. How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, pages 177 – 190.

Evert, Stefan. 2009. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook, article 58*. Mouton de Gruyter, Berlin.

Fernando, Chitra and Roger Flavell. 1981. *On Idiom: Critical Views and Perspectives*. Exeter Linguistic Studies 5. University of Exeter.

Fillmore, Charles J. 1968. The case for case. In Emmon Bach and Robert Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York, pages 1–88.

Firth, John Rupert. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society, Oxford, pages 1–32.

Francis, Gill. 1993. A corpus-driven approach to grammar – principles, methods and examples. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology*. Benjamins, Amsterdam, pages 137–156.

Frank, Robert. 2002. *Phrase Structure Composition and Syntactic Dependencies*. MIT Press, Cambridge, MA.

Fries, Udo. 2010. Sentence length, sentence complexity and the noun phrase in the 18th-century news publication. In M. Kytö, J. Scahill, and H. Tanabe, editors, *Language Change and Variation from Old English to Late Modern English: A Festschrift for Minoji Akimoto*. Peter Lang, Bern, pages 21–34.

Fundel, K., R. Küffner, and R. Zimmer. 2007. RelEx – relation extraction extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Gale, William, Kenneth Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the ARPA Workshop on Speech and Natural Language Processing*, pages 233 – 237.

Gamon, Michael. 2011. High-order sequence modeling for language learner error detection. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 180–189, Portland, Oregon, June. Association for Computational Linguistics.

Giuliano, Claudio, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of EACL 2006*, pages 401–408.

Grefenstette, Edward, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. 2011. Concrete sentence spaces for compositional distributional models of meaning. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, Oxford.

Gries, Stefan Th. 2010. Methodological skills in corpus linguistics: a polemic and some pointers towards quantitative methods. In Tony Harris and María Moreno Jaén, editors, *Corpus linguistics in language teaching*. Peter Lang, Frankfurt a.M., pages 121–146.

Gries, Stefan Th. 2012. Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: towards more and more fruitful exchanges. In Joybrato Mukherjee and Magnus Huber, editors, *Corpus linguistics and variation in English: Theory and description*. Rodopi, Amsterdam, pages 41–63.

Gries, Stefan Th. and Stefanie Wulff. 2005. Do foreign language learners also have constructions? evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics*, 3:182–200.

Gries, Stefan Th. and Stefanie Wulff. 2009. Psycholinguistic and corpus linguistic evidence for l2 constructions. *Annual Review of Cognitive Linguistics*, 7:163–186.

Grover, Claire. 2008. LT-TTT2 example pipelines documentation. Technical report, Edinburgh Language Technology Group,.

Harris, Zellig. 1968. *Mathematical Structures of Language*. Wiley, New York.

Harris, Zellig. 1970. Distributional structure. In *Papers in structural and transformational Linguistics*. pages 775–794.

Haverinen, Katri, Filip Ginter, Sampo Pyysalo, and Tapio Salakoski. 2008. Accurate conversion of dependency parses: targeting the Stanford scheme. In Tapio Salakoski, Dietrich Rebholz-Schuhmann, and Sampo Pyysalo, editors, *Proceedings of Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, pages 133–136, Turku, Finland. Turku Centre for Computer Science (TUCS).

Hirschman, L, A Yeh, C Blaschke, and A Valencia. 2005. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1.

Hoey, Michael. 2005. *Lexical priming: A New Theory of Words and Language*. Routledge.

Hundt, Marianne, David Denison, and Gerold Schneider. 2012. Retrieving relatives from historical data. *Literary and Linguistic Computing*, 27(1):3–16.

Hundt, Marianne, David Dension, and Gerold Schneider. 2012. Relative complexity in scientific discourse. *English Language and Linguistics*, 16(2):209–240.

Hunston, Susan and Gill Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Benjamins, Amsterdam/Philadelphia.

Ishikawa, Shin. 2009. Vocabulary in interlanguage: A study on corpus of english essays written by asian university students (ceeaus). In K. Yagi and T. Kanzaki, editors, *Phraseology, corpus linguistics and lexicography: Papers from Phraseology 2009 in Japan*, pages 87–100, Nishinomiya, Japan. Kwansei Gakuin University Press.

Jucker, Andreas H. 1993. The genitive versus the of-construction in newspaper language. In Andreas H. Jucker, editor, *The Noun Phrase in English: Its Structure and Variability*. Universitätsverlag Winter, Heidelberg, pages 121–136.

Jucker, Andreas H., Gerold Schneider, Irma Taavitsainen, and Barb Breustedt. 2007. Fishing for compliments: Precision and recall in corpus-linguistic compliment research. In Andreas H. Jucker and Irma Taavitsainen, editors, *Speech Act History of English (Pragmatics & Beyond New Series)*. John Benjamins, Amsterdam/Philadelphia.

Kaljurand, Kaarel, Gerold Schneider, and Fabio Rinaldi. 2009. UZurich in the BioNLP 2009 Shared Task. In *Proceedings of the BioNLP workshop, Boulder, Colorado*.

Kaplan, Ronald M., John T. Maxwell III, Tracy Holloway King, and Richard S. Crouch. 2004. Integrating finite-state technology with deep LFG grammars. In *ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP (ComShaDeP 2004)*, Nancy, France.

Keller, Frank. 2003. A probabilistic parser as a model of global processing difficulty. In Richard Alterman and David Kirsh, editors, *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 646–651, Boston.

Keller, Frank. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 317–324, Barcelona.

Keller, Frank. 2010. Cognitively plausible models of human language processing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 60–67.

Kim, J.D., T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - a semantically annotated corpus for biotextmining. *Bioinformatics*, 19(1):180–182.

Kim, S., J. Yoon, and J. Yang. 2008. Kernel approaches for genic interaction extraction. *Bioinformatics*, 9:10.

Krallinger, M., F. Leitner, C. Rodriguez-Penagos, and A. Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of biocreative ii. *Genome Biology*, 9(Suppl 2).

Krallinger, Martin, Miguel Vazquez, Florian Leitner, David Salgado, Andrew Chatr-aryamontri, Andrew Winter, Livia Perfetto, Leonardo Briganti, Luana Licata, Marta Iannuccelli, Luisa Castagnoli, Gianni Cesareni, Mike Tyers, Gerold Schneider, Fabio Rinaldi, Robert Leaman, Graciela Gonzalez, Sergio Matos, Sun Kim, W Wilbur, Luis Rocha, Hagit Shatkay, Ashish Tendulkar, Shashank Agarwal, Feifan Liu, Xinglong Wang, Rafal Rak, Keith Noto, Charles Elkan, Zhiyong Lu, Rezarta Dogan, Jean-Fred Fontaine, Miguel Andrade-Navarro, and Alfonso Valencia. 2011. The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics*, 12(Suppl 8):S3.

Kreyer, Rolf. 2003. Genitive and of-construction in modern written english: Processability and human involvement. *International Journal of Corpus Linguistics*, 8(2):169–207.

Kreyer, Rolf. 2010. *Introduction to English Syntax*. Text-books in English Language and Linguistics. Peter Lang, Frankfurt a. M.

Krug, Manfred. 2003. Frequency as a determinant in grammatical variation and change. In Rohdenburg and Mondorf (Rohdenburg and Mondorf, 2003), pages 7–67.

Labov, William. 1969. Contraction, deletion, and inherent variability of the english copula. *Language*, 45(4):715–762.

Leacock, Claudia, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*. Number 25 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, second edition edition.

Lee, Daniel D. and H. Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, pages 556–562.

Lee, David Y. W. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. *Language Learning and Technology*, Vol.5(3):37–72.

Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith. 2009. *Change in Contemporary English. A Grammatical Study*. Cambridge University Press, Cambridge.

Lehmann, Hans Martin and Gerold Schneider. 2009. Parser-based analysis of syntax-lexis interaction. In Andreas H. Jucker, Daniel Schreier, and Marianne Hundt, editors, *Corpora: Pragmatics and discourse: papers from the 29th International conference on English language research on computerized corpora (ICAME 29)*, Language and computers 68. Rodopi, Amsterdam/Atlanta, pages 477–502.

Lehmann, Hans Martin and Gerold Schneider. 2011. A large-scale investigation of verb-attached prepositional phrases. In S. Hoffmann, P. Rayson, and G. Leech, editors, *Studies in Variation, Contacts and Change in English, Volume 6: Methodological and Historical Dimensions of Corpus Linguistics*. Varieng, Helsinki.

Lehmann, Hans Martin and Gerold Schneider. 2012a. Dependency bank. In *Proceedings of LREC 2012 Workshop on Challenges in the management of large corpora*, pages 23–28.

Lehmann, Hans Martin and Gerold Schneider. 2012b. Syntactic variation and lexical preference in the dative-shift alternation. In Joybrato Mukherjee and Magnus Huber, editors, *Studies in Variation, Contacts and Change in English, Papers from the 31st International conference on English language research on computerized corpora (ICAME 31), Giessen, Germany*. Rodopi, Amsterdam.

Leitner, Florian, Scott A. Mardis, Martin Krallinger, Gianni Cesareni, Lynette A. Hirschman, and Alfonso Valencia. 2010. An overview of biocreative ii.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):385–399.

Levin, Beth C. 1993. *English Verb Classes and Alternations: a Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Levy, Roger and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

Lin, Dekang. 1999. Automatic identification of noncompositional phrases. In Steven Bird, editor, *Proceedings of the 37th Annual Meeting of the ACL*, pages 317–324, College Park, USA. Association for Computational Linguistics (ACL).

López-Couso, Maria José, Bas Aarts, and Belén Méndez-Naya. 2012. Late modern english syntax. In Alexander Bergs and Laurel J. Brinton, editors, *Historical Linguistics of English: An international handbook. Volume I*, Handbooks of Linguistics and Communication Science [HSK] 34.1. Mouton de Gruyter, pages 869–887.

MacDonald, M. C. and M. S. Seidenberg. 2006. Constraint satisfaction accounts of lexical and sentence comprehension. In M. J. Traxler and M. A. Gernsbacher, editors, *Handbook of psycholinguistics*. Elsevier, London, 2nd ed. edition, pages 581–611.

Mair, Christian. 2009. Corpus linguistics meets sociolinguistics: the role of corpus evidence in the study of sociolinguistic variation and change. In Antoinette Renouf and Andrew Kehoe, editors, *Corpus Linguistics: Refinements and Reassessments*. Rodopi, Amsterdam, pages 7–32.

Malvern, David D., Brian J. Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical Diversity and Language Development*. Palgrave MacMillan, Houndmills, UK.

Manes, Joan and Nessa Wolfson. 1981. The compliment formula. In Florian Coulmas, editor, *Conversational Routine. Explorations in Standardized Communication Situations and Prepatterned Speech*. Mouton, The Hague, pages 115 – 132.

Marcus, Mitch, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

Martin, Bronwen and Felizitas Ringham. 2000. *Dictionary of Semiotics*. Cassell, New York.

McCarthy, Diana, Bill Keller and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition, and Treatment*, pages 73–80. Association for Computational Linguistics (ACL).

Meyers, Adam, Catherine Macleod, Roman Yangarber, Ralph

Grishman, Leslie Barrett, and Ruth Reeves. 1998. Using NOMLEX to produce nominalization patterns for information extraction. In *Coling-ACL98 workshop Proceedings: the Computational Treatment of Nominals*, Montreal, Canada.

Mikheev, Andrei. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics*, 23(3):405–423.

Millar, Neil. 2011. The processing of malformed learner collocations. *Applied Linguistics*, 32(2):129–148.

Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6:1–28.

Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford University Press, Oxford.

Moon, Rosamund. 2009. *Words, grammar, text: revisiting the work of John Sinclair*. Benjamins.

Mukherjee, Joybrato. 2004. Corpus data in a usage-based cognitive gramma. In Karin Aijmer and Bengt Altenberg, editors, *Corpus Data in a Usage-based Cognitive Grammar*. Rodopi, Amsterdam, pages 85–100.

Mukherjee, Joybrato. 2005. *English Ditransitive Verbs: Aspects of Theory, Description and a Usage-based Model*. Language and Computers, ed. Christian Mair, Charles F. Meyer and Nelleke Oostdijk. Rodopi, Amsterdam/New York.

Mukherjee, Joybrato and Sebastian Hoffmann. 2006. Describing verb-complementational profiles of new englishes: A pilot study of indian english. *English World-Wide*, 27(2):147–173.

Nelson, Gerald. 2003. Modals of obligation and necessity in varieties of english. In Pam Peters, editor, *From Local to Global English*. Dictionary Research Centre, Macquarie University, Sydney, pages 25–32.

Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Varieties of English Around the World: G29. John Benjamins, Amsterdam.

Ninio, Anat, 2006. *Language and the Learning Curve: A new theory of syntactic development*, chapter Lexicalism. Oxford University Press, Oxford.

Nivre, Joakim. 2006. *Inductive Dependency Parsing*. Text, Speech and Language Technology 34. Springer, Dordrecht, The Netherlands.

Pawley, Andrew and Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. C. Richards and R. W. Schmidt, editors, *Language and Communication*. Longman, London, pages 191–226.

Pyysalo, S., F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski. 2007. Bioinfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics, 8(50).*

Ratnaparkhi, Adwait. 1996. A Maximum Entropy Part-Of-Speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania.

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the bard: Evaluating the accuracy of a modern pos tagger on early modern english corpora. In *Proceedings of Corpus Linguistics 2007*. University of Birmingham, UK.

Rebholz-Schuhmann, D., H.Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P.Stoehr. 2006. EBIMed – text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23(2):e237 – e244.

Rinaldi, Fabio, Simon Clematide, and Gerold Schneider. 2010. Ontogene in calbc. In *Proceedings of the CALBC workshop.*

Rinaldi, Fabio, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. 2008. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13.

Rinaldi, Fabio, Gerold Schneider, and Simon Clematide. 2012. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics*.

Rinaldi, Fabio, Gerold Schneider, Simon Clematide, and Gintare Grigonyte. 2012. Notes about the ontogene pipeline. In *AAAI-2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*, Arlington, Virginia, USA.

Rinaldi, Fabio, Gerold Schneider, Simon Clematide, Silvan Jegen, Pierre Parisot, Martin Romacker, and Therese Vachon. 2010a. Ontogene (team 65): preliminary analysis of participation in biocreative iii. In *Proceedings of BioCreative III*, Bethesda, Maryland, USA., September 13-15.

Rinaldi, Fabio, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Thérèse Vachon, and Martin Romacker. 2010b. Ontogene in biocreative ii.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 3:472–480.

Rinaldi, Fabio, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. 2006. An environment for relation mining over richly annotated corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3):S3.

Rohdenburg, Günter and Britta Mondorf, editors. 2003. *Determinants of Grammatical Variation in English*. Topics in English Linguistics 43. Mouton de Gruyter, Berlin and New York.

Ronan, Patricia and Gerold Schneider. submitted. Investigating light verb constructions in contemporary british and irish english. In *Paper presented at ICAME 2013*.

Rosenbach, Anette. 2002. *Genitive variation in English. Conceptual factors in synchronic and diachronic studies*. Mouton de Gruyter, Berlin.

Rothenhäusler, Klaus and Hinrich Schütze. 2009. Unsupervised classification with dependency based word spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 17–24, Athens, Greece, March. Association for Computational Linguistics.

Röthlisberger, Melanie and Gerold Schneider. 2013. Of-genitive versus s-genitive: A corpus-based analysis of possessive constructions in 20thcentury English. In Paul Bennet, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpora*, Korpuslinguistik und Interdisziplinäre Perspektiven auf Sprache - Corpus linguistics and Interdisciplinary perspectives on language (CLIP). Narr Francke Attempto, Stuttgart.

Sacks, Harvey. 1995. *Lectures on Conversation, Vol. 1*. Blackwell, Oxford.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multi-word expressions: A pain in the neck for nlp. Technical Report LinGO Working Paper No. 2001-03, Stanford University, CA.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Ph.D. thesis, Stockholm University.

Sand, Andrea. 2004. Shared morpho-syntactic features in contact varieties of English: Article use. *World Englishes*, 23:281–98.

Sangkuhl, Katrin, Dorit S. Berlin, Russ B. Altman, and Teri E. Klein. 2008. PharmGKB: Understanding the effects of individual genetic variants. *Drug Metabolism Reviews*, 40(4):539–551. PMID: 18949600.

Sankoff, David. 1988. Sociolinguistics and syntactic variation. In F.J. Newmeyer, editor, *Linguistics: the Cambridge Survey*. Cambridge University Press, Cambridge, pages 140–61.

Schmid, Hans-Jörg. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Mouton de Gruyter, Berlin.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Schneider, Edgar. 2004. How to trace structural nativization: Particle verbs in World Englishes. *World Englishes*, 23:2:227–249.

Schneider, Gerold. 2005. A broad-coverage, representationally minimal LFG parser: chunks and F-structures are sufficient. In Mriram Butt and Traci Holloway King, editors, *The 10th international LFG Conference (LFG 2005)*, Bergen, Norway. CSLI.

Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Schneider, Gerold. 2012a. Adapting a parser to historical English. In Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen, and Matti Rissanen, editors, *Studies in Variation, Contacts and Change in English, Volume 10: Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*, Helsinki, Finland.

Schneider, Gerold. 2012b. Using semantic resources to improve a syntactic dependency parser. In Viktor Pekar Verginica Barbu Mititelu, Octavian Popescu, editor, *SEM-II workshop at LREC 2012*.

Schneider, Gerold. 2013. Using automatically parsed corpora to discover lexico-grammatical features of English varieties. In Fryni Kakoyianni Doa, editor, *Penser le Lexique-Grammaire, perspectives actuelles*, Colloques, Congrès et Conférences – Sciences du Langage, Histoire de la Langue et des Dictionnaires. Éditions Honoré Champion, Paris, pages 491–504.

Schneider, Gerold. accepted for publication. Automated media content analysis from the perspective of computational linguistics. In *Inhaltsanalysen und Neue Medien*. Herbert Halem.

Schneider, Gerold, Simon Clematide, Gintare Grigonyte, and Fabio Rinaldi. 2012. Using syntax features and document discourse for relation extraction on PharmGKB and CTD. In Fabio Rinaldi, editor, *Proceedings of SMBM 2012*, Zurich.

Schneider, Gerold, Simon Clematide, and Fabio Rinaldi. 2011. Detection of interaction articles and experimental methods in biomedical literature. *BMC Bioinformatics*, special issue on BioCreative III.

Schneider, Gerold and Gintare Grigonyte. 2013. Using an automatic parser as a language learner model. In *Book of Abstracts of LCR 2013*, Bergen, Norway, September.

Schneider, Gerold and Marianne Hundt. 2009. Using a parser as a heuristic tool for the description of New Englishes. In *Proceedings of Corpus Linguistics 2009*, Liverpool.

Schneider, Gerold and Marianne Hundt. 2012. “Off with their heads” – profiling TAM in ICE corpora. In Marianne Hundt and Ulrike Gut, editors, *Mapping Univity and Diversity world-wide*, VEAW. Benjamins, Amsterdam.

Schneider, Gerold, Kaarel Kaljurand, Thomas Kappeler, and Fabio Rinaldi. 2009. Detecting protein-protein interactions in biomedical texts using a parser and linguistic resources. In *Proceedings of CICLING 2009*.

Schneider, Gerold, Kaarel Kaljurand, and Fabio Rinaldi. 2009. Detecting Protein/Protein Interactions using a parser and linguistic resources. In *CICLing 2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 406–417, Mexico City, Mexico. Springer LNC 5449.

Schneider, Gerold, Hans Martin Lehmann, and Peter Schneider. 2014. Parsing Early Modern English corpora. *Literary and Linguistic Computing*, first published online February 6, 2014 doi:10.1093/llc/fqu001.

Schneider, Gerold and Fabio Rinaldi. 2011. A data-driven approach to alternations based on protein-protein interactions. In *Proceedings of the 3rd Congreso Internacional de Lingüística de Corpus (CILC), Valencia, Spain, 7-9 April, 2011*.

Schneider, Gerold, Fabio Rinaldi, Kaarel Kaljurand, and Michael Hess. 2005. Closing the gap: Cognitively adequate, fast broad-coverage grammatical role parsing. In *ICEIS Workshop on Natural Language Understanding and Cognitive Science (NLUCS 2005)*, Miami, FL, May 2005.

Schneider, Gerold and Heinrich Zimmermann. 2010. Text Mining Methoden im Semantic Web. *Praxis der Wirtschaftsinformatik*, HMD(271):36–47.

Schneider, Gerold and Lena Zipp. 2013. Discovering new verb-preposition combinations in New Englishes. In Joybrato Mukherjee and Magnus Huber, editors, *Studies in Variation, Contacts and Change in English, Volume 14 – Corpus Linguistics and Variation in English: Focus on non-native Englishes*. Varieng, Helsinki.

Schone, Patrick and Dan Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In Lillian Lee and Donna Harman, editors, *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing*, pages 100–108, Pittsburgh, Pennsylvania. Association for Computational Linguistics (ACL).

Schönefeld, Doris. 1999. Corpus linguistics and cognitivism. *International Journal of Corpus Linguistics*, 4:131–171.

Schulte im Walde, Sabine and Alissa Melinger. 2008. An in-depth look into the co-occurrence distribution of semantic associates. *Italian Journal of Linguistics. Special Issue on From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science*, 20(1):89–128.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Schwartz, AS and MA Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput*, pages 451–462.

Sedlatschek, Andreas. 2009. *Contemporary Indian English: variation and change*. Varieties of English around the world. John Benjamins, Amsterdam / Philadelphia.

Seoane, Elena. 2009. Syntactic complexity, discourse status and animacy as determinants of grammatical variation in modern english. *English Language and Linguistics*, 13(3):365–384.

Seretan, Violeta. 2011. *Syntax-Based Collocation Extraction*. Springer, Dordrecht.

Seretan, Violeta and Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 953–960, Sydney, Australia, July. Association for Computational Linguistics.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. OUP, Oxford.

Sinclair, John McHardy and Ronald Carter. 2004. *Trust the Text: Language, Corpus And Discourse*. Routledge.

Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, pages 209–43.

Stubbs, Michael. 1995. Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language*, 2(1):23–55.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English. A corpus studyat the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Mouton de Gruyter, Berlin and New York.

Terkourafi, Marina. 2001. *Politeness in Cypriot Greek: A frame-based approach*. Ph.D. thesis, Department of Linguistics, University of Cambridge.

Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Librairie Klincksieck, Paris.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. John Benjams, Amsterdam.

Tomasello, Michael. 2000. The item based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4:156–163.

Tuggener, Don, Manfred Klenner, Gerold Schneider, Simon Clematide, and Fabio Rinaldi. 2011. An incremental model for the coreference resolution task of BioNLP 2011. In *Proceedings of the BioNLP11 shared task. Portland, Oregon, 24 June, 2011*.

van Noord, Gertjan and Gosse Bouma. 2009. Parsed corpora for linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 33–39, Athens, Greece. Association for Computational Linguistics.

Volk, Martin and Gerold Schneider. 1998. Comparing a statistical and a rule-based tagger for german. In *Proceedings of KONVENS-98*, pages 125–137, Bonn.

Wasow, Thomas and Jennifer Arnold. 2003. Post-verbal constituent ordering in english. In Rohdenburg and Mondorf (Rohdenburg and Mondorf, 2003).

Watts, Richard. 2003. *Politeness*. Cambridge University Press, Cambridge.

Weeds, Julie, James Dowdall, Gerold Schneider, Bill Keller, and David Weir. 2007. Using distributional similarity to organise BioMedical terminology. In Fidelia Ibekwe-SanJuan, Anne Condamines, and M. Teresa Cabré Castellví, editors, *Application-Driven Terminology Engineering*. Benjamins, Amsterdam/Philadelphia.

Wiegers, T, A Davis, KB Cohen, L Hirschman, and C Mattingly. 2009. Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (ctd). *BMC Bioinformatics*, 10(1):326.

Wueest, Bruno, Gerold Schneider, and Michael Amsler. 2014. Measuring the public accountability of new modes of governance. In *ACL Workshop on Language Technologies and Computational Social Science*, Baltimore, Maryland, June 26.

Wulff, Stefanie. 2008. *Rethinking Idiomaticity*. Research in Corpus and Discourse. Continuum, London.

Yule, George U. 1944. *The statistical study of literary vocabulary*. Cambridge University Press, Cambridge.

Zipf, George Kingsley. 1965. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. The MIT Press, Cambridge, Massachusetts.