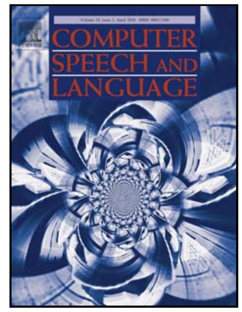
Geography of Social Ontologies: Testing a Variant of the Sapir-Whorf Hypothesis in the Context of Wikipedia

Alexander Mehler, Olga Pustylnikov, Nils Diewald

Accepted Manuscript



Title: Geography of Social Ontologies: Testing a Variant of the Sapir-Whorf Hypothesis in the Context of Wikipedia

Authors: Alexander Mehler, Olga Pustylnikov, Nils Diewald

PII: S0885-2308(10)00043-4
DOI: doi:10.1016/j.csl.2010.05.006
Reference: YCSLA 459

To appear in:

Received date: 31-10-2009
Revised date: 10-3-2010
Accepted date: 7-5-2010

Please cite this article as: Mehler, A., Pustylnikov, O., Diewald, N., Geography of Social Ontologies: Testing a Variant of the Sapir-Whorf Hypothesis in the Context of Wikipedia, *Computer Speech & Language* (2008), doi:10.1016/j.csl.2010.05.006

Geography of Social Ontologies: Testing a Variant of the Sapir-Whorf Hypothesis in the Context of Wikipedia

Alexander Mehler[a], Olga Pustylnikov[a], Nils Diewald[a]

[a] *Text Technology, Faculty of Technology, Bielefeld University, Universitätsstraße 25, D-33615 Bielefeld, Germany*

Abstract

In this article, we test a variant of the Sapir-Whorf Hypothesis in the area of complex network theory. This is done by analyzing social ontologies as a new resource for automatic language classification. Our method is to solely explore structural features of social ontologies in order to predict family resemblances of languages used by the corresponding communities to build these ontologies. This approach is based on a reformulation of the Sapir-Whorf Hypothesis in terms of distributed cognition. Starting from a corpus of 160 Wikipedia-based social ontologies, we test our variant of the Sapir-Whorf Hypothesis by several experiments, and find out that we outperform the corresponding baselines. All in all, the article develops an approach to classify linguistic networks of tens of thousands of vertices by exploring a small range of mathematically well-established topological indices.

Key words: Sapir-Whorf Hypothesis, linguistic networks, automatic language classification, social ontologies, quantitative network analysis

1. Introduction

This article presents an approach to automatic language classification based on complex network theory [1–3]. It explores the topologies of social ontologies as part of Wikipedia to get a new data source of genealogical classification. In so doing, the article tests a variant of the *Sapir-Whorf Hypothesis* (SWH) by means of a network-theoretical approach. It tackles the question, whether structural similarities of social ontologies correspond to family resemblances of the underlying languages.

Generally speaking, the SWH states that language structure imprints on cognitive structure [4, 5]. If this principle of linguistic relativity is true, then the usage of similar languages should result in similar conceptual structures.

Email addresses: Alexander.Mehler@uni-bielefeld.de (Alexander Mehler), Olga.Pustylnikov@uni-bielefeld.de (Olga Pustylnikov), Nils.Diewald@uni-bielefeld.de (Nils Diewald)

Therefore, conversely, conceptual structures should be indicative of family resemblances of the languages in which they are manifested. According to our network-theoretical approach, we additionally hypothesize that these resemblances can be deduced from topological similarities of conceptual structures.

To test the SWH, we explore *conceptual structures* in terms of *social ontologies* as a sort of linguistic network in which vertices denote terminological units while edges stand for terminological relations of subordination [6]. This approach is indispensable as social ontologies are to date the only access point to large scale conceptual structures in numerous languages of various families.[1] As these systems are based on terminological relations of subordination according to the wiki principle [7], we speak of social ontologies as instances of social tagging [8], which extend the range of terminological ontologies [6].

Note that we do not use the notion of a social ontology in terms of philosophy [9], nor in the sense that a social ontology is an ontology of social entities. In contrast to this, the attribute 'social' relates to the wiki principle by which the ontologies under consideration are generated. That is, social ontologies as manifested by the category systems of Wikipedia are non-automatically, manually generated by their users according to guidelines[2] which may vary between different languages.

Generally speaking, a social ontology emerges as a solution to a coordination problem among large groups of interacting agents [10]. This relates to the sharing of a collaboratively-structured, dynamically-growing universe of semantic units [11]. Social ontologies as exemplified by the category system of Wikipedia [12] manifest the output of a kind of distributed cognition [13], which is distributed among agents who collaboratively generate and structure certain fields of knowledge. By utilizing social ontologies as a resource of language classification, we specify the general notion of cognition in the formulation of the SWH as that of *distributed* cognition. As a consequence, we arrive at a variant of the SWH, which states that *language structure imprints on distributed cognition as manifested by social ontologies so that their topologies are indicative of the corresponding language families.* We present a series of experiments to test this hypothesis.

The article is organized as follows: Section 2 brings our variant of the SWH in line with research on this hypothesis. Related approaches to language classification are discussed in Section 3. Further, Section 4 informs about the corpus of social ontologies explored in this study. Our method to formalize these ontologies, to quantify their topology, and to automatically classify them is presented in Section 5. Based on that, Section 6 tests our target hypotheses and discusses our findings. Finally, Section 7 gives a conclusion.

[1] As of October 2009, there are Wikipedias for 271 languages, each of which includes a category system that manifests conceptual structures shared by the underlying community of wikilocutors (see `http://meta.wikimedia.org/wiki/List_of_Wikipedias/sortable`).

[2] See, for example, `http://de.wikipedia.org/wiki/Wikipedia:Kategorien` in relation to `http://en.wikipedia.org/wiki/Wikipedia:Categorization`.

2. Towards a Variant of the Sapir-Whorf Hypothesis

Few ideas have caused as much controversy and debate in linguistics as the Sapir-Whorf Hypothesis. Basically, it states that language influences the way in which we think about reality [4]. One reason for the controversy about this hypothesis relates to its variant in terms of the principle of *linguistic determinism*, which implies, for example, the impossibility of translations. Notwithstanding this disputable variant, there is a less controversial version in terms of the principle of *linguistic relativity*. This version claims that language influences thought by acting as a mediator between reality and its conceptual representation [14]. Despite this common understanding of language as a mediator, approaches to the SWH are distinguished by their perspective on this role [14]:

- *Structure-centered* approaches start from an observed structural difference between languages (e.g., on the level of single linguistic constructions). They refer to this variation as the *explanandum* and try to explain it by means of differences in the experience of reality and its conceptual representation (the *explanans*). One problem with this approach is its uncritical selection of particular languages as quasi-neutral reference points for comparison. Whorf's classical comparison of the verbalization of time in Hopi and English falls into this class of approaches. As a matter of fact, his study is heavily disputed in linguistics [15, see 16 for a discussion].
- *Domain-centered* approaches focus on a selected domain of experience (*explanandum*) (e.g., the range of colors [5, 17]) to ask how particular languages structure this domain (*explanans*). Unlike structure-centered approaches, the scope of investigation of the linguistic anticipation of the domain is narrow. In any event, this approach makes it possible to precisely compare large numbers of languages [18]. However, comparisons of this sort are biased by the small range of categories under consideration (e.g., color terms [19]) and the selection of the domain-related terms according to linguistic introspection [14].
- Finally, *behavior-centered* approaches try to explain behavioral differences (*explanandum*) by linguistic differences (*explanans*). Obviously, these approaches reverse the perspective of their structure-centered counterparts. An example is Whorf's [4] observation of how different readings of the word 'empty' caused accidental fires. In any event, this approach is biased by the difficulty of verifying the salience and strength of the relation between linguistic features and the observed behavior [20–23].

The present study combines the *domain-centered* with the *structure-centered* approach. On the one hand, our method can be regarded as *domain-centered* since we refer to encyclopedic domains as the data source of language classification. At the same time, we overcome the restriction of traditional domain-centered approaches to small ranges of terms. The reason is that social ontologies cover, in principle, the complete range of encyclopedic knowledge and its terminological manifestation. Additionally, we circumvent the problematic introspection of many domain-centered approaches as we access social ontologies directly without any subjective mediation. Consequently, we depart from domain-centered approaches in two respects. The first is that we do not compare the terms of different ontologies directly, nor do we directly compare the referents of these terms in the corresponding domains. Rather, we follow a strict network-theoretical approach as outlined in Section 1.

This approach is inspired by experiments that demonstrate the expressiveness of exclusively structural classifications of linguistic units [24]. Dimter [25], for example, asked subjects to guess the type of texts (e.g., weather forecast, obituary announcement etc.) in which all content words had been replaced by random strings. Surprisingly, most test persons guessed these types correctly, obviously *by exploring the structure of the texts*.[3] In this article, we transfer Dimter's approach to the level of linguistic networks: we explore topologies of networks in contrast to text structures in order to classify language families instead of text types. In this sense, our approach is *structure-centered* in that we ask, *how* wikilocutors organize encyclopedic domains depending on the languages that they use. Starting with social ontologies, we look at structural differences of their topologies and ask whether wikilocutors of related languages organize encyclopedic domains in a similar way.

Altogether, this combines to a *domain-structure-centered* approach since we ask whether wikilocutors of related languages structure encyclopedic domains in a similar way. As our variant of the SWH focuses on *distributed* cognition, our approach cannot directly be compared with recent findings on neural correlates of the SWH [19], which focus on the cognition of *single* agents. However, as we enlarge the scope of the SWH, this may help to bridge these two areas of cognitive science. Our approach directly relates to a variant of the SWH that has been recently formulated by Nisbett [28].

2.1. Nisbett's Hypothesis

In his book "Geography of Thought" [28], Richard E. Nisbett compares the Western tradition based on the philosophy of Ancient Greek to the Eastern tradition shaped by several other philosophies [28, pp. 12]. He argues that differences in these cultural traditions – as manifested in language – have different influences on the speakers' behaviors. For example, Nisbett observes that Indo-European languages all have expressions for abstract nouns, whereas Chinese does not (e.g., there is no direct translation for 'size' in Chinese, nor does this language have a suffix '-ness' with which to build abstract nouns).

[3] In a pretest, we successfully automatized Dimter's experiment by classifying more than 30,000 texts into 31 text classes [26, 27]. This has been done by accentuating the structure-oriented stance of Dimter's experiment. In our trials, we deleted any content words so that the classifier had no information about the length of the words, nor about numbers and their text position – actually, this information was retained in Dimter's experiment. The only information used by our algorithm was the logical document structure of input texts (in terms of the hierarchical nesting of sections, captions, paragraphs and sentences) while it disregarded all lexical information (except from the number of tokens). Based on this information, more than 70% of the texts were classified correctly.

Nisbett reports a wide range of psycholinguistic experiments in support of his hypothesis. Recent results from other studies concerned with the East-West comparison are referred to in [20, 21, 23, 29].[4] The present study tests Nisbett's Hypothesis by means of exploring social ontologies. That is, different social ontologies belonging to a particular cultural group – Western or Eastern – are tested for similarities within the group and in contrast to each other.

In summary, we test two related hypotheses: a variant of the SWH and a variant of the closely related hypothesis of Nisbett. Note that the latter variant is a special case of our variant of the SWH as it focuses on the manifestation of cultural differences in terms of the topologies of social ontologies.

3. Related Work to Language Classification

Generally speaking, language classification aims to categorize languages by means of their genealogical descent. The basic idea is that languages inherit structural features from a common root so that they can be ascribed to the same family. By measuring different degrees of similarity between languages, a language family tree can be reconstructed (often referred to as glossogeny [31]).

Early lexicostatistical approaches – closely connected to the name of Morris Swadesh – were solely based on calculating differences between the lexical material of pairs of languages [32]. The main units of these approaches are so-called *cognates*. These are pairs of words taken from different languages that have the same meaning, coincide in their phonetic/phonological form, and originate from a common ancestor in a (hypothetical) parental language. The degree of relatedness of two languages is then calculated by the number of shared cognates which occur in a limited list of pairs of *core* words that are synonymous in both languages. In the beginning, the decision of whether or not two words were phonetically similar was made based on intuition [32, 33]. Subsequently, algorithms were developed to formalize these judgments [e.g., 34, 35]. In many cases, these algorithms used the character-based edit distance of words [36, 37], sometimes enhanced by phonetic criteria [38, see 39 for a survey on phonetic string matching].

According to Swadesh, the list of core word forms is the primary access point to what he calls the fundamental vocabulary of a language, which supposedly covers the part of the lexicon that is mostly independent from cultural influences. Following an assumption made by Sapir, that "[the] greater the degrees of linguistic differentiation within a stock, the greater is the period of time that must be assumed for the development of such differentiations" [40, p.76], Swadesh [32] proposes that this vocabulary changes at a roughly constant rate over time. This hypothesis is the starting point for the reconstruction of language family trees complemented by information about the probable time of language divergence (in the style of carbon-14 dating in archaeology). Even

[4] Of course, this differentiation in behavior does not imply any difference in cognitive abilities [29, 30].

though Swadesh's universal glottochronological approach was quickly disputed [41, 42], it led to a better understanding of language change and motivated further studies on its variation rate. For example, [43] show a significant variation based on the frequency of word use: the more frequently a word is used in a language, the slower it evolves over time.

The existence of loanwords is another effect that takes part in lexicostatistics and influences the variation of change rates. Besides their inheritance from a common origin, languages can share cognates by borrowing in areal neighborhood. While Swadesh reduces the borrowability of words to the non-fundamental, *cultural* part of vocabularies (and sorts them out of his core lists), [44] argue that the borrowability of a word (or a grammatical construction [45]) depends on its frequency of use similar to the variation of its change rate.[5] Recent models additionally account for such geographical effects [47].

Despite the success of glossogenetic reconstructions by lexicostatistics, the validity of inter-lexical comparison for language classification and family tree reconstruction is controversial. This is not only due to the lack of additional linguistic features (such as morphological or syntactical aspects), but also in respect to a debated incomparability of phonetic forms throughout languages: Cognates must be objectively transcribed into a common phonetic space and it is highly questionable whether such a common space exists or not [see 48 for a discussion].

Consequently, newer approaches concentrate on intra- rather than inter-language comparisons. These approaches generate profiles of languages in order to calculate their dissimilarity. They compare, for example, confusion probability matrices (as a kind of intra-language edit-distance matrix) [44], n-gram profiles [49], or typological feature vectors [50]. Finally, approaches to network-based language profiles calculate dissimilarities of languages by means of topological differences. This relates, for example, to explorations of phoneme networks [51] or so-called *Global Syntactic Dependency Networks* (GSDNs) [2, 52].

A central advantage of the network-theoretical approach as followed here is that it disposes of direct comparisons of lexical units or typological features. Rather, it opens the door to topological information as a novel resource for language classification. Thus, with our approach to language classification based on comparisons of the structures of ontologies, we aim to avoid known shortcomings of lexicostatistics with a simple, yet comprehensive model.

4. A Corpus of Social Ontologies

In order to study our variant of the SWH and its descendant in the form of Nisbett's Hypothesis, we explore a corpus of social ontologies from Wikipedia, which is henceforth called *Social Ontology Corpus* (SOC). Table 1 and 2 show

[5] However, recent investigations question a direct correlation between stability and borrowability of words [46].

Table 1: Wikimedia codes of 160 Wikipedias (underlined) whose social ontologies have been analyzed in this article. They have been selected because their largest weakly connected component contains at least 100 vertices (see `http://meta.wikimedia.org/wiki/List_of_Wikipedias/sortable`).

aa ab <u>af</u> ak <u>als</u> <u>am</u> <u>an</u> <u>ang</u> <u>ar</u> arc as <u>ast</u> av ay <u>az</u> ba <u>bar</u> <u>bat-smg</u> bcl <u>be</u> <u>be-x-old</u>
<u>bg</u> bh bi bm <u>bn</u> bo <u>bpy</u> <u>br</u> <u>bs</u> bug bxr <u>ca</u> <u>cbk-zam</u> cdo ce <u>ceb</u> ch cho chr chy <u>co</u> cr <u>crh</u>
<u>cs</u> <u>csb</u> <u>cu</u> <u>cv</u> <u>cy</u> <u>da</u> <u>de</u> diq <u>dsb</u> dv dz ee <u>el</u> eml en <u>eo</u> <u>es</u> <u>et</u> <u>eu</u> <u>ext</u> <u>fa</u> ff <u>fi</u> <u>fiu-vro</u>
fj <u>fo</u> <u>fr</u> <u>frp</u> <u>fur</u> <u>fy</u> <u>ga</u> <u>gan</u> <u>gd</u> <u>gl</u> glk <u>gn</u> got gu <u>gv</u> ha hak haw <u>he</u> <u>hi</u> hif ho <u>hr</u> <u>hsb</u>
<u>ht</u> <u>hu</u> <u>hy</u> <u>hz</u> <u>ia</u> <u>id</u> <u>ie</u> ig ii ik ilo <u>io</u> <u>is</u> <u>it</u> iu <u>ja</u> <u>jbo</u> <u>jv</u> <u>ka</u> <u>kaa</u> kab kg ki kj <u>kk</u> kl
<u>km</u> <u>kn</u> <u>ko</u> kr ks <u>ksh</u> <u>ku</u> <u>kv</u> <u>kw</u> ky <u>la</u> <u>lad</u> <u>lb</u> lbe lg <u>li</u> <u>lij</u> <u>lmo</u> <u>ln</u> <u>lo</u> <u>lt</u> <u>lv</u> map-bms <u>mdf</u>
mg mh <u>mi</u> <u>mk</u> <u>ml</u> <u>mn</u> <u>mo</u> <u>mr</u> <u>ms</u> <u>mt</u> mus my <u>myv</u> mzn <u>na</u> <u>nah</u> <u>nap</u> <u>nds</u> <u>nds-nl</u> <u>ne</u> <u>new</u> ng <u>nl</u> <u>nn</u>
<u>no</u> nov nrm nv ny <u>oc</u> om or <u>os</u> pa pag <u>pam</u> pap <u>pdc</u> pi <u>pih</u> <u>pl</u> <u>pms</u> ps <u>pt</u> <u>qu</u> <u>rm</u> <u>rmy</u> rn
<u>ro</u> roa-rup <u>roa-tara</u> <u>ru</u> rw sa sah sc <u>scn</u> <u>sco</u> sd <u>se</u> sg <u>sh</u> <u>si</u> <u>simple</u> <u>sk</u> <u>sl</u> sm sn so
<u>sq</u> <u>sr</u> <u>srn</u> ss st stq <u>su</u> <u>sv</u> <u>sw</u> <u>szl</u> <u>ta</u> <u>te</u> tet <u>tg</u> <u>th</u> ti tk <u>tl</u> tn <u>to</u> tokipona tpi <u>tr</u> ts
<u>tt</u> tum tw ty <u>udm</u> ug <u>uk</u> <u>ur</u> <u>uz</u> ve <u>vec</u> <u>vi</u> <u>vls</u> <u>vo</u> <u>wa</u> <u>war</u> <u>wo</u> <u>wuu</u> xal xh <u>yi</u> yo za <u>zea</u> <u>zh</u>
<u>zh-classical</u> <u>zh-min-nan</u> <u>zh-yue</u> zu

the Eurasian-centered distribution of the releases of Wikipedia that have been analyzed here. This corpus has been analyzed in two ways:

- The corpus of 160 social ontologies (see Table 1) of at least 100 vertices in their largest weakly connected component has been analyzed in order to study the separability of various topological indices (see Section 5.2). This has been done to select those indices which best separate the different ontologies *only by virtue of their topology*. The ontologies in this corpus range from a minimum order of 103 vertices and a minimal size of 102 arcs to a maximum order of 102,129 vertices and a maximum order of 205,391 arcs (see Table 3). These 160 ontologies have on average 8,348.9 vertices (order) and 14,634 arcs (size).[6] To the best of our knowledge, this is the largest corpus of social ontologies that has been analyzed so far.[7]
- Based on this corpus, we have selected several subcorpora of Western and Eastern languages in order to perform experiments in genealogical language classification according to our variant of the SWH. With the exception of the English Wikipedia, the elements of these subcorpora have been selected according to their size: for a given language family, ontologies were selected whose order is of at least 1,000 vertices. See Table 2 for a complete listing of the experiments based on these subcorpora.

5. A Network Model of Ontology-Based Language Classification

Our variant of the SWH states that the structure of social ontologies is indicative of family resemblances of the underlying languages. In this section,

[6] The data was downloaded in November and December, 2008.

[7] It can be downloaded from `www.linguistic-networks.net` (Resources/Corpora/Social Software). Note that we have transformed all ontologies into GraphML [53] in order to secure the text-technological sustainability of our social ontology corpus.

Table 2: The list of 46 social ontologies considered in 7 experiments E0–E6 (Section 6) on language classification including the pilot study E0. The table reports the Wikimedia codes of the ontologies together with the names of the corresponding languages, their mapping onto language families as well as the order (#vertices) and size (#arcs) of the ontologies. Finally, columns E0–E6 report which languages have been considered in which experiment.

code	name	family	area	order	size	E0	E1	E2	E3	E4	E5	E6
zh	Chinese	Sinitic	Eastern 1	38,468	68,903					×	×	×
zh-classical	Classical Chinese	Sinitic	Eastern 1	1,115	1,123					×	×	×
zh-yue	Cantonese	Sinitic	Eastern 1	3,839	5,214					×	×	×
ja	Japanese	Japonic	Eastern 1	54,362	115,713					×	×	×
ko	Korean	Korean	Eastern 1	28,708	53,174						×	×
id	Indonesian	Sundic	Eastern 2	25,781	43,137							×
ms	Malay	Sundic	Eastern 2	4,922	7,915							×
su	Sundanese	Sundic	Eastern 2	4,365	5,050							×
af	Afrikaans	Germanic	Western	2,262	3,248				×	×	×	×
da	Danish	Germanic	Western	13,727	23,542	×	×	×	×	×	×	×
de	German	Germanic	Western	58,466	114,421	×	×	×	×	×	×	×
fy	West Frisian	Germanic	Western	1,609	1,949				×	×	×	×
is	Icelandic	Germanic	Western	9,344	13,964			×	×	×	×	×
ksh	Ripuarian	Germanic	Western	2,245	4,635				×	×	×	×
lb	Luxembourgish	Germanic	Western	6,892	10,463			×	×	×	×	×
nds	Low German	Germanic	Western	1,620	2,142				×	×	×	×
nl	Dutch	Germanic	Western	37,192	69,505	×	×	×	×	×	×	×
nn	Norwegian Nynorsk	Germanic	Western	13,928	25,605			×	×	×	×	×
no	Norwegian	Germanic	Western	25,984	45,457			×	×	×	×	×
sv	Swedish	Germanic	Western	40,777	72,996	×	×	×	×	×	×	×
an	Aragonese	Romanic	Western	4,901	6,585				×	×	×	×
ast	Asturian	Romanic	Western	2,362	3,016				×	×	×	×
ca	Catalan	Romanic	Western	11,556	19,729	×	×	×	×	×	×	×
es	Spanish	Romanic	Western	68,471	126,633	×	×	×	×	×	×	×
fr	French	Romanic	Western	102,129	205,391			×	×	×	×	×
gl	Galician	Romanic	Western	4,540	5,929				×	×	×	×
it	Italian	Romanic	Western	59,259	107,473	×	×	×	×	×	×	×
la	Latin	Romanic	Western	5,274	7,394			×	×	×	×	×
oc	Occitan	Romanic	Western	7,049	13,128			×	×	×	×	×
pms	Piedmontese	Romanic	Western	1,548	1,834				×	×	×	×
pt	Portuguese	Romanic	Western	48,229	100,986			×	×	×	×	×
ro	Romanian	Romanic	Western	28,513	49,060	×	×	×	×	×	×	×
be	Belarusian	Slavic	Western	4,449	5,414				×	×	×	×
be-x-old	Belarusian Taraškievica	Slavic	Western	17,118	36,438			×	×	×	×	×
bg	Bulgarian	Slavic	Western	8,453	15,213	×	×	×	×	×	×	×
bs	Bosnian	Slavic	Western	15,220	21,301			×	×	×	×	×
cs	Czech	Slavic	Western	24,830	44,295	×	×	×	×	×	×	×
hr	Croatian	Slavic	Western	7,207	12,524			×	×	×	×	×
mk	Macedonian	Slavic	Western	10,999	19,146			×	×	×	×	×
pl	Polish	Slavic	Western	37,796	62,434			×	×	×	×	×
ru	Russian	Slavic	Western	63,772	118,871	×	×	×	×	×	×	×
sh	Serbo-Croatian	Slavic	Western	2,364	3,087				×	×	×	×
sk	Slovak	Slavic	Western	24,730	43,200			×	×	×	×	×
sl	Slovenian	Slavic	Western	24,526	44,785	×	×	×	×	×	×	×
sr	Serbian	Slavic	Western	11,941	16,743			×	×	×	×	×
uk	Ukrainian	Slavic	Western	17,781	30,557			×	×	×	×	×
AVG / SUM				21,535	39,333	12	12	28	38	42	43	46

we make this hypothesis a measurable and testable property. This is done in four steps:

Step 1. *by specifying the formal class of graphs spanned by social ontologies;*
Step 2. *by identifying topological characteristics of this class of graphs*;
Step 3. *by representing social ontologies by vectors of these topological indices*;
Step 4. *by using these feature vectors as input to automatic classification.*

The first two steps relate to our representation model of social ontologies and are explained in Section 5.1 and 5.2, respectively. The last two steps are covered by cluster analysis. As the features in use denote topological indices of networks, we subsume these two steps under the notion of *quantitative network analysis* (see Section 5.3).

Table 3: Some statistical characteristics of the *Social Ontology Corpus* (SOC) analyzed here: *height* is the eccentricity [54] of the main category v of the corresponding social ontology, *width* is the maximum number of vertices with equal distance to the root and *level* is the corresponding distance for which this maximum is reached.

	order	**size**	**height**	**width**	**level**
minimum	103	102	2	20	1
median	1,570	1,949	9	582	4
maximum	102,129	205,391	30	34,181	13
average	8,348.9	14,634	9.8616	2,774.9	4.5031
standard deviation	15,377	29,583	4.1269	5,169	2.1784

5.1. Social Ontologies as Directed Acyclic Graphs

Figure 1 exemplifies the kind of relations which form the skeleton of social ontologies. It shows an outline of the category system of the English Wikipedia in which the category *mammal* is subordinated to the categories *vertebrates*, *tetrapods*, and *synapsids*, while it subordinates, for example, the categories *bats*, *primates* and *fur*. This example shows that in social ontologies, subordination does not necessarily coincide with hyperonymy relations (*fur* is not a kind of *mammal*). Figure 1 also shows that social ontologies may include cycles: in the German Wikipedia, the category *Druckerzeugnis* [print product] is subordinated to *Buch* [book], which is subordinated to *Bibliothekswesen* [librarianship] which is finally subordinated to *Druckerzeugnis*. Figure 1 also shows a larger cyclic structure as part of the social ontology of the Turkish Wikipedia. Obviously, ontologies of this sort do *not* span trees, but a certain class of more general graphs as exemplified in Figure 2. It shows three Wikipedia-based category systems of approximately the same size. As exemplified by these digraphs, social ontologies do not span trees, but graphs with a kernel hierarchical structure that is superimposed by arcs which add a graph-like structure. Figure 2 also hints at the fact that the widths of social ontologies grow more than their depths. This is shown in Figure 3 (left), which reports the ratio of depth and width of the 160 ontologies in our SOC. Obviously, for a growing order (i.e., number of vertices) this ratio is close to zero.

Figure 4 presents a schematic account of this class of graphs. From left to right we observe an increase in structural complexity: while graph (b) generalizes tree (a) by graph-inducing downward, upward and lateral arcs, graph (c) additionally possesses a second source [56] called A. It is this third graph that best captures the scenario of social ontologies, which may contain multiple sources, cycles and even loops. However, social ontologies are neither trees, nor acyclic or arbitrary graphs. Rather, they form a class in the range of these extreme cases: graphs which are spanned around a kernel hierarchy that build *nearly* acyclic graphs [57]. That is, social ontologies contain cycles, but not very many. This is shown in Figure 3 (right), which reports the number of vertices that belong to cycles in relation to the order of the ontologies in our SOC. Obviously, this ratio is mostly near but not equal to zero.

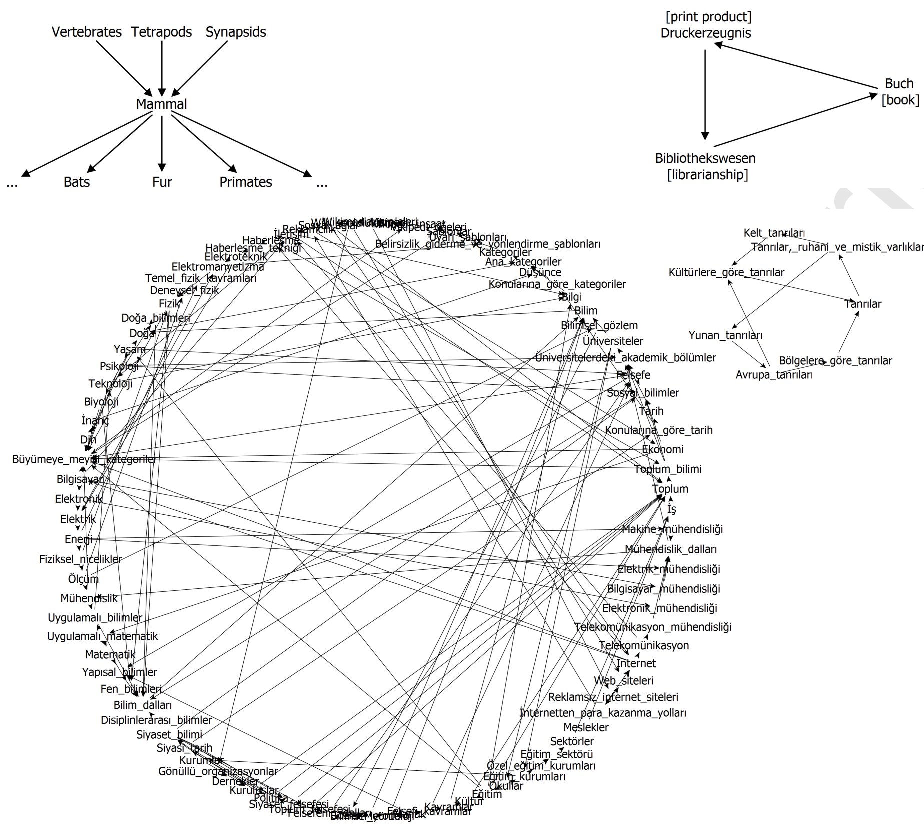


Figure 1: *Left (top):* Hyperonyms and hyponyms from the point of view of the category 'Mammal' in the category graph of the English Wikipedia. Arcs go from the superordinate category to its subordinate. *Right (top):* a cyclic structure of three categories in the social ontology of the German Wikipedia. *Bottom:* Cyclic structures in the social ontology of the Turkish Wikipedia.

To give a formal definition of this class of graphs we extend the notion of a directed generalized tree [55, 58], which is, in turn, based on the notion of a tree. The reason to proceed in this way is that while *Directed Acyclic Graphs* (DAG) generalize the notion of a tree, social ontologies have a graph-like structure which extends the one of generalized trees as they are spanned around a kernel DAG-like structure. It is necessary to consider this class of graphs in formal terms as it constrains the set of network indices that actually characterize social ontologies. Definition 1 and 2 provide this formal account.

Definition 1. Let $T = (V, A', r)$ be a directed tree rooted in $r \in V$. Further, for any vertex $v \in V$ let $P_{rv} = (v_{i_0}, a_{j_1}, v_{i_1}, \ldots, v_{i_{n-1}}, a_{j_n}, v_{i_n}), v_{i_0} = r, v_{i_n} = v, a_{j_k} \in A', in(a_{j_k}) = v_{i_{k-1}}, out(a_{j_k}) = v_{i_k}, 1 \leq k \leq n$, be the unique path in T from r to v such that $V(P_{rv}) = \{v_{i_0}, \ldots, v_{i_n}\}$ is the set of all vertices on that path. A *Directed Generalized Tree* $G = (V, A_{1..5}, r)$ based on the kernel

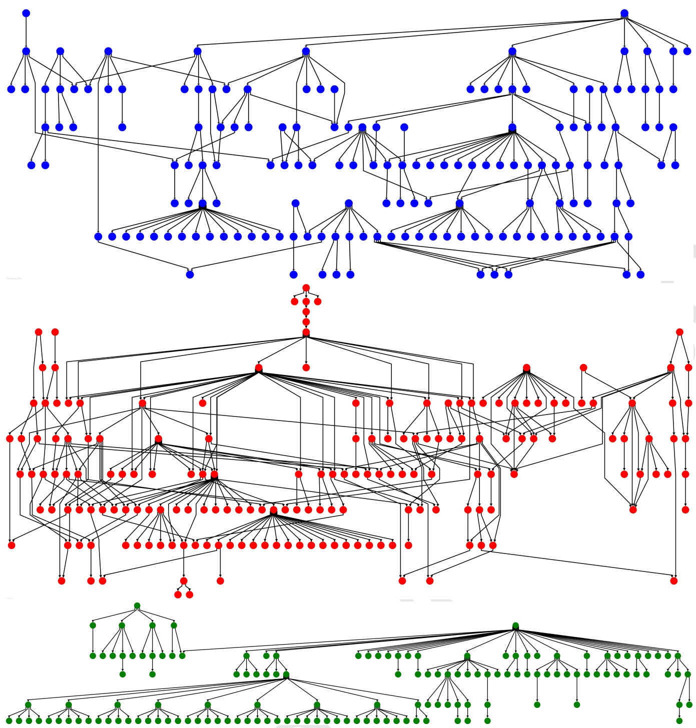


Figure 2: The largest connected component of the category system of the *Friulian* Wikipedia (A), the *Northern Sami* Wikipedia (B) and the *Moksha* Wikipedia (C) – Moksha is an Uralic language spoken in Mordovia. It belongs, with Northern Sami, to the Finno-permic languages.

tree T is a pseudograph whose arc set is partitioned so that $A_{1..5} = \cup_{i=1}^{5} A_i$, $\forall 1 \leq i < j \leq 5\colon A_i \cap A_j = \emptyset$ and $a \in A_{1..5}$ iff $a \in \cup_{i=1}^{5} A_i$ and

$$\begin{array}{lcll}
a \in A_1 & = & A' & \textit{(kernel arcs)} \\
a \in A_2 & \subseteq & \{a \,|\, in(a) = v \in V \wedge out(a) = w \in V(P_{rv}) \setminus \{v\}\} & \textit{(upward arcs)} \\
a \in A_3 & \subseteq & \{a \,|\, in(a) = w \in V(P_{rv}) \setminus \{v\} \;\wedge\; out(a) = v \in V\} & \textit{(downward arcs)} \\
a \in A_4 & \subseteq & \{a \,|\, in(a) = out(a) \in V\} & \textit{(reflexive arcs)} \\
a \in A_5 & \subseteq & V^2 \setminus (A_1 \cup A_2 \cup A_3 \cup A_4) & \textit{(lateral arcs)}
\end{array}$$

G is said to be generalized by its reflexive, lateral, up- and downward arcs.

Graph (b) in Figure 4 exemplifies a generalized tree. Graphs of this sort are quite common in web-based communication [59]. They provide a blueprint for defining generalized nearly acyclic graphs (see Figure 4.C) that naturally extend generalized trees in the sense of the following definition.

Definition 2. Let $G' = (V, A', S)$ be a *Directed Acyclic Graph* (DAG) with the set of sources $S \subseteq V$ and $\mathbb{P}(G')$ be the set of all paths in G' such that $\forall r \in S \forall v \in V : |\{(x, \ldots, y) \in \mathbb{P}(G') \,|\, x = r \wedge y = v\}| \leq 1$. We denote this unique path (that excludes the existence of downward arcs) by $P_{rv} \in \mathbb{P}(G')$, if

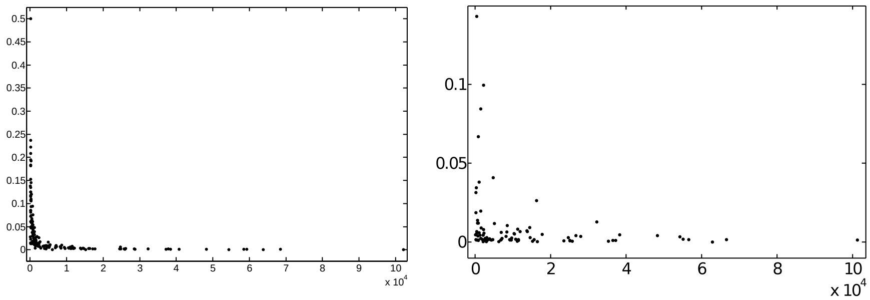


Figure 3: *Left:* the ratio of depth and width of the largest connected component (y-axis) in relation to the order (x-axis) of 160 social ontologies in our SOC. *Right:* the ratio $C/|V|$ of the number C of vertices that belong to cycles and the order $|V|$ of the ontologies (y-axis) as a function of $|V|$ (x-axis).

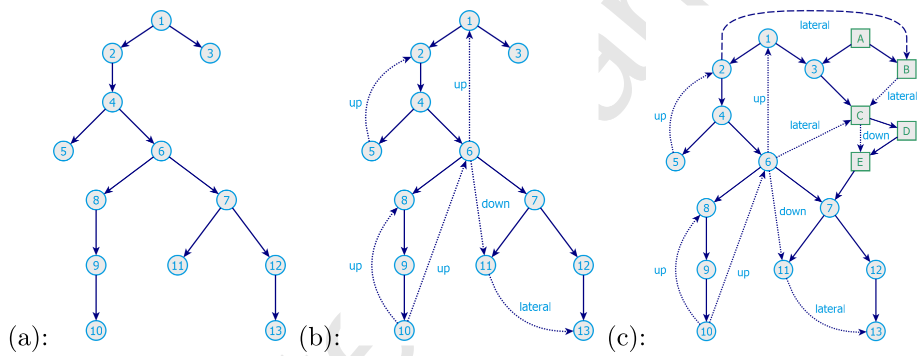


Figure 4: (a): A directed tree rooted by vertex 1. (b): a generalized directed tree with the same kernel hierarchical structure in conjunction with four upward arcs, one downward and one lateral arc [55]. (c): a structural scenario that resembles social ontologies, that is, a graph with two sources, 1 and A, whose kernel (resulting from deleting all upward and lateral arcs) spans a directly acyclic graph.

it exists, and write $P_{rv} \notin \mathbb{P}(G')$, if not. Additionally, we demand that G' does not contain lateral arcs that connect vertices reachable from different sources: $\forall r, s \in S : (r \neq s \wedge P_{rv}, P_{sw} \in \mathbb{P}(G') \wedge P_{rw}, P_{sv} \notin \mathbb{P}(G')) \Rightarrow \neg\exists(v, \ldots, w) \in \mathbb{P}(G')$. A *Generalized Acyclic Graph* $G = (V, A_{1..5}, S)$ based on the DAG G' is a graph such that $A_{1..5} = \cup_{i=1}^{5} A_i$, $\forall 1 \leq i < j \leq 5 \colon A_i \cap A_j = \emptyset$ and[8]

$$\begin{array}{lcl}
a \in A_1 & = & a \in A' \\
a \in A_2 & \subseteq & \{a \,|\, \exists r \in S \exists P_{rv} \in \mathbb{P}(G') \colon in(a) = v \in V \wedge out(a) = w \in V(P_{rv}) \setminus \{v\}\} \\
a \in A_3 & \subseteq & \{a \,|\, \exists r \in S \exists P_{rv} \in \mathbb{P}(G') \colon in(a) = w \in V(P_{rv}) \setminus \{v\} \;\wedge\; out(a) = v \in V\} \\
a \in A_4 & \subseteq & \{a \,|\, in(a) = out(a) \in V\} \\
a \in A_5 & \subseteq & V^2 \setminus (A_1 \cup A_2 \cup A_3 \cup A_4)
\end{array}$$

[8]Note that, as defined in Definition 1, $V(P) \subseteq V$ is the set of vertices on the path P.

G is called a *Generalized Nearly Acyclic Graph* (GNAG) if its number C of vertices that enter into cycles is small in relation to its order, that is, if $0 < C/|V| \ll 1$.

Obviously, Figure 3 (right) shows that social ontologies are indeed characteristic in that their number of vertices that enter into cycles is close (but not necessarily equal) to 0.

In order to capture the structure of social ontologies according to this graph model we need to go beyond network theory, which deals with less restricted graphs. In short, we may explore social ontologies as networks because of their cyclicity. However, because of their kernel hierarchy, we may explore them as acyclic graphs or even as trees. This plurality is captured by our quantitative model of social ontologies.

5.2. Topological Fingerprints of Directed Acyclic Graphs

In this section we present our approach to characterizing social ontologies by topological indices of their graph model. As explained in the last section, we capture both the network- and tree-like structures of social ontologies in a single model. This is done by taking fingerprints of GNAGs by means of four classes of topological indices:

Class 1. *Network Theoretical (NT) measures:* We utilize the apparatus of scale-free networks [1]. In a pilot study (see Section 6.1), we test the hypothesis that languages can be classified into families based on topological indices of dependency networks as invented by [2]. In line with this approach, we test whether the same indices indicate the membership of social ontologies to language families. We test this for the cluster coefficients C_{ws} [60], C_{br} [61] and their weighted counterparts $\langle C_w(k)\rangle$ and $\langle C_w^{ns}(k)\rangle$ [62]. Further, we consider the diameter δ together with the average geodesic distance $\langle L\rangle$, the average degree, Newman's assortativity index [1] and the expected $\langle L\rangle$ and C_{ws} of the random and regular graphs of equal order and size.[9] All in all, we consider 12 indices in Class 1 – see [57] for a thorough exemplification of these indices in the context of linguistic networks.

Class 2. *Information Theoretical (IT) Measures:* In addition, we investigate a range of measures that have been invented in order to describe the information content of graphs and processes of information flow based on them [see 54 for a first introduction into this topic]. This relates to so-called measures of graph entropy [64]. The idea behind this approach is more related to Nisbett's Hypothesis, which states that information

[9] As GNAGs are more restricted than general networks, the exponent of the power law that best fits to the out-degree distribution of vertices [63] together with its adjusted coefficient of determination [57] do not make sense as indices here.

content tells us something about the shareability [65] of knowledge systems. Therefore, we direct our attention to this class of topological indices. Further, a pre-study has shown that compactness and centrality measures are informative about differences of linguistic networks like such as wiki graphs [57]. This includes the compactness measure of hypertext theory [66] as well as graph-related centrality measures such as graph, degree and closeness centrality, which have been successfully applied in NLP [67]. As centrality measures are primarily based on the notion of geodesic distance, they relate to graph entropy measures so that we commonly refer to this group as *Information Theoretical* (IT) measures. All in all, we experiment with 45 indices in Class 2 as further described in Section 5.2.1.

Class 3. *GNAG-based Measures:* We additionally utilize a range of measures that have been developed in order to capture the topological specifics of social ontologies in contrast to terminological and formal ontologies [68]. This class of measures is sensitive to the kernel hierarchical structure of GNAGs and, therefore, goes beyond network-theoretical indices (of Class 1). We experiment with 52 indices in Class 3 as described in Section 5.2.2.

Class 4. *Measures related to a Sensitivity Analysis (SA):* As a fourth class of features, rather than beginning a new measurement, we instead undertake a deterministic selection among all $109 = 12 + 45 + 52$ topological indices described so far. That is, we compute for each index I how well it differentiates among all 160 social ontologies in our SOC (see Section 4). This is done by means of the *sensitivity measure* $S(I)$ of Konstantinova et al. [69] for a topological index I:

$$S(I) = \frac{|C| - |C_i|}{|C|} \in [0, 1] \quad (1)$$

where C_i is the set of networks from the SOC C that I cannot distinguish. These are networks for which there is at least one other network in C that is mapped onto the same number by I. As we know that all ontologies in C are pairwise different, we ask whether a candidate topological index accounts for this difference by mapping the networks onto different numbers. Indices I for which $S(I) \rightarrow 0$ are called *degenerated* [69, 70]. The results of computing S for our indices can be seen in Figure 5. It shows that 34 of 109 indices distinguish exactly 100% of the networks correctly. These 34 indices are collected in Class 4 as they are minimally degenerated in terms of S.[10] As an alternative to this subset, we consider the set of indices that are degenerated only by 5%. These are indices, which distinguish at least 95% of the networks in our reference SOC.

[10]Interestingly, C_{ws}, C_{br} and diameter, for example, are deselected in this way.

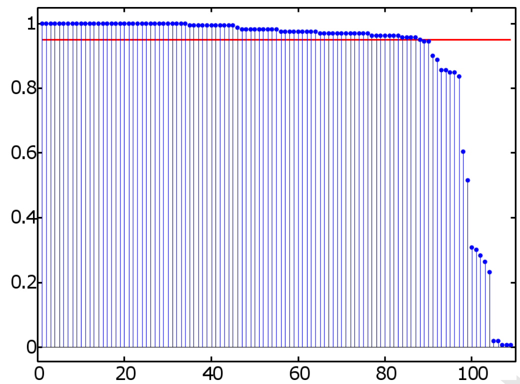


Figure 5: Sensitivity measure of 109 indices based on the approach of [69]. The horizontal line denotes the 95% limit.

The measures that fall into the first class have been extensively discussed in the literature [see, e.g., 1, 71–73 and 67 for thorough introductions]. In this article, we concentrate on a short presentation of measures in Class 2 and 3.

5.2.1. Graph Entropy

The literature discusses a wide range of measures of graph entropy [64]. One approach to apply the notion of entropy H to a vertex $v \in V$ in a graph $G = (V, E)$ is to say that $H(v)$ codes information about the topology of G from the perspective of v: if the geodesic distances from v to the other vertices of G are uniformly distributed, $H(v)$ is high. In this case, we are little determined in entering the neighborhood of v, when randomly selecting v as an entry point to G. An extreme case is a star graph around the center v. In case of a social ontology this is tantamount to a very flat, but broad ontology. If, in contrast to this, the distances are non-uniformly distributed, so that $H(v)$ is low, we are more determined in traversing the neighborhood of v, when selecting v as our starting point. Now, an extremal case would be a line graph starting from v. In case of a social ontology this is tantamount to a very deep, but narrow ontology. Dehmer [74] has generalized the vertex-related entropy to arrive at a class of entropy measures of graphs – we use its variant from [70]:

$$H_{f_{\mathbf{c}}}(G) = -\sum_{v \in V} \frac{f_{\mathbf{c}}(v)}{\sum_{w \in V} f_{\mathbf{c}}(w)} \log \left(\frac{f_{\mathbf{c}}(v)}{\sum_{w \in V} f_{\mathbf{c}}(w)} \right) \tag{2}$$

where

$$f_{\mathbf{c}}(v) = \sum_{j=1}^{\delta(G)} c_j |S_j(v)| \quad ; \quad S_j(v) = \{w \in V \mid \delta(v, w) = j\} \tag{3}$$

$\delta(v, w)$ is the geodesic distance of v and w in G, $\mathbf{c}' = (c_1, \ldots, c_{\delta(G)})$ is a vector of weights $c_i \geq 0$, $\sum_i c_i > 0$, used to bias the so-called j-spheres S_j, and

$\delta(G)$ is the diameter of G. By varying $\mathbf{c}$, we get different instances of the class of entropy measures in Equation 2. In this article, we experiment with exponentially and logistically decaying weights. This is done in order to simulate processes of growth and of disintegration of spreading activation [75]. In this way, we obtain a scheme for experimenting with entropy-related measures where a genetic algorithm is used to finally select those measures that are most characteristic of ontologies. All in all, we experiment with 45 entropy, centrality and compactness-related measures as elements of Class 2 indices [see 57, 74 for thorough discussions of them].

5.2.2. Imbalance

Social ontologies have been contrasted with classification schemes (e.g., the DDC), terminological ontologies (e.g., WordNet), and formal ontologies (e.g., the *Suggested Upper Merged Ontology*) [6]. Using a small range of topological indices, the membership of an ontology to one of these classes has been correctly predicted in 93% of the ontologies considered [68]. As these indices are indicative of different *types* of ontologies, they may also be indicative of different *families* of *social* ontologies. We aim to test this assumption by utilizing the approach of [68], which basically explores the imbalance of the graphs spanned by ontologies as follows: let $D = (V, A)$ be a directed graph and $x \in V$ be a distinguished vertex (e.g., the one that denotes the main category in the ontology represented by D). Further, let $Q\colon S_1(x) \to [0,1]$ be an interval-scaled function such that $\forall v_{i_j} \in \{v_{i_1}, \ldots, v_{i_n}\} = S_1(x)\colon Q(v_{i_j}) \geq 0$ and $\sum_{j=1}^{n} Q(v_{i_j}) = 1$ so that we get a feature vector

$$\mathbf{q}(x) = (Q(x_{i_1}), \ldots, Q(x_{i_n}))' = (q_1, \ldots, q_n)' \tag{4}$$

as input to the relative entropy to measure the *balance* of D from the point of view of x with respect to Q:

$$RH(\mathbf{q}(x)) = \frac{H(\mathbf{q}(x))}{\log_2 n} = -\frac{\sum_{i=1}^{n} q_i \log_2 q_i}{\log_2 n} \in [0,1] \tag{5}$$

Finally, we define a *measure of imbalance* I_Q of x in D induced by Q by means of the redundancy measure R [76]:

$$I_Q(x) = R(\mathbf{q}(x)) = 1 - RH(\mathbf{q}(x)) \in [0,1] \tag{6}$$

Equation 6 gives a scheme for measuring the imbalance of a digraph D from the point of view of the distinguished vertex x according to the attribute Q. By varying this attribute we get alternative measures of imbalance of D. Following [68], we consider the *depth*, *width*, *level*, *order*, *length* (the number of leafs within the scope of the focal node), *complexity* (the number of immediate constituents) and *dependency* (as a function of the number of vertices subordinated in a tree-like structure [77]) as different attributes. Taking the main category as the distinguished vertex, these attributes allow for characterizing ontologies with respect to their intricacy of design, richness of detail and related structural attributes. We experiment with 52 such Class 3 indices [see 68 for a thorough discussion of them]. We expect that social ontologies that belong to different language families are quite distinguishable by these attributes as they reflect the specifics of the class of graphs instantiated by these ontologies, that is, GNAGs.

5.3. Quantitative Network Analysis

Using the structural information captured by topological indices of social ontologies, we can classify this sort of networks by means of cluster analysis. More specifically, we apply *Quantitative Network Analysis* (QNA) [57, 68] in order to learn classes of social ontologies *by virtue of their structure*, while disregarding any content units (i.e., names of vertices). QNA basically integrates vector representations of complex networks with hierarchical cluster analysis. The cluster analysis is complemented by a subsequent partitioning, where the number of classes is determined in advance. In this sense, QNA is semi-supervised. We experiment with *single*, *complete*, *average*, and *weighted* linkage, while we use the *Mahalanobis distance*, the *(standardized) Euclidean distance* and two distance measures based on Pearson's *correlation coefficient* and on the *cosine measure*, respectively, to compute pairwise object distances.

Roughly speaking, QNA takes the space of input objects together with the parameter space of linkage methods and distance measures to find out the parameter constellation, which best separates the data in terms of the corresponding gold standard [see 57 for a thorough explanation of this approach]. Note that we use F-measure statistics (i.e., the harmonic mean of precision and recall) to evaluate our classification results.[11] Note also that QNA integrates a genetic search of the best performing subset of topological indices that maximizes the F-score of the corresponding classification. As a matter of fact, this search tries to find the optimal feature set, but may also stop at a local maximum. In order to handle correlations between different indices and to scale down the parameter space, we use the Mahalanobis distance whenever possible.

6. Experimentation

6.1. A Pilot Study: Network-Based Classification of Languages

In this section, we present a pilot study to automatically classify languages into genealogical groups based on syntactic networks. We do that to get insights into the possibilities of network-based language classification in general. This pilot study serves as a linguistically well-motivated basis of comparison to evaluate the outcomes of our social-ontology-based approach. In order to provide this basis of comparison, we use a syntactic resource to generate the networks in this pre-test. This relates to so-called *Global Syntactic Dependency Networks* (GSDN) as introduced by Ferrer i Cancho et al. [78]. In graph theoretical terms,

[11] Precision and recall are computed with respect to a gold standard which in our case is the partition of the set of languages into language families. F ranges in the interval $[0, 1]$. 1 indicates a perfect and 0 the worst classification.

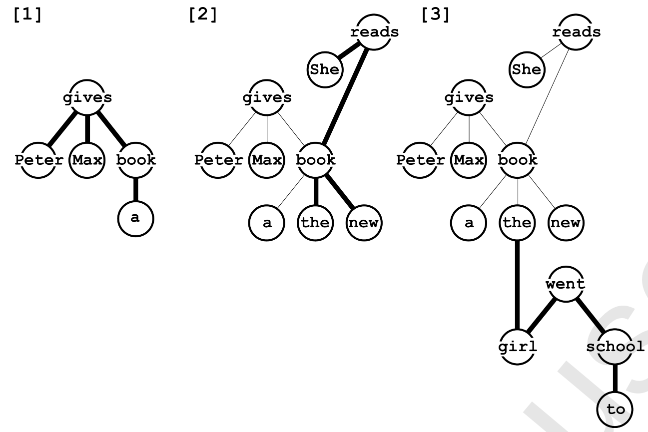


Figure 6: Three steps in creating a GSDN by processing the first three sentences in a dependency treebank.

GSDNs are undirected networks with multiple edges. *Vertices* of a GSDN represent word forms of a dependency treebank[12], while *edges* represent syntactic dependency relations. A GSDN of a particular language is constructed from its corresponding treebank as exemplified in Figure 6. The input treebank is parsed sentence by sentence so that word forms are added as vertices to the target network. Vertices are inserted only once.[13] If in subsequent sentences a vertex (word) appears again as part of a new dependency relation, more edges are added to it (e.g., *book* in Figure 6).

We construct GSDNs from 12 dependency treebanks as listed in Table 4. In order to test whether GSDNs reflect genealogical differences of languages, we represent them by means of a subset of 24 of the 109 features (see Section 5.2) and make the resulting vectors an input to QNA (see Section 5.3). Our aim is to classify the vectors into three genetic groups (i.e., *Slavic*, *Germanic*, and *Romanic*). In addition, we apply two baseline scenarios to evaluate the goodness of our results. Both scenarios randomly assign languages to one of the three groups. The *known-partition*-scenario has knowledge about the cardinality of each target class, whereas the *equi-partition*-scenario assumes an equal size of each group. The computation of the baselines is repeated 1,000 times so that finally their average F-scores are considered.

The results of the pilot study are presented in Table 5. Surprisingly, we get a maximum F-score of 1, which is produced by using only 8 features as a result of applying a genetic search for the best performing subset of features. Figure 7 shows the corresponding dendrogram. This result indicates a high potential of

[12] A *dependency treebank* is a corpus in which each sentence is annotated regarding its syntactic dependency structure [79].

[13] Note that multiple edges are represented by edge weights, which denote frequencies of co-occurrence.

Table 4: The 12 treebanks that have been used to generate GSDNs in the pilot study.

Treebank	Language	$\|V\|$	$\|E\|$	Reference
Alpino Treebank v.1.2	Dutch	28,475	102,184	[80]
Danish Dependency Treebank v.1.0	Danish	19,133	50,858	[81]
A sample of sentences of the Dependency Grammar Annotator	Romanian	8,867	23,901	[82]
Russian National Corpus	Russian	58,283	177,942	[83]
A sample of the Slovene Dependency Treebank v.0.4	Slovene	8,342	20,453	[84]
Talkbanken05 v.1.1	Swedish	25,097	126,526	[85]
Turin University Treebank v.0.1	Italian	7,984	24,269	[86]
Catalan Dependency Treebank (CESS)	Catalan	38,882	215,308	[87]
Spanish Dependency Treebank (Cast3LB)	Spanish	17,101	56,911	[88]
Prague Dependency Treebank 2.0	Czech	146,504	696,379	[89]
BulTreeBank	Bulgarian	32,421	95,698	[90]
Tiger Treebank	German	2,465	4,399	[91]

language classification by means of GSDNs. However, if we take all 24 features into account, the corresponding F-score falls down to 63%, which is still above the corresponding baseline of around 55%. Obviously, some of the features in this set of topological indices bias the classification. On the other hand, a small range of only 8 indices suffices to separate the languages correctly. This subset includes, amongst others, the cluster coefficient of Watts and Strogatz [60], the γ of the power law of the corresponding degree distribution and the centrality measures considered here.[14]

The dendrogram in Figure 7 shows that although languages are grouped correctly into clusters, the similarities within the cluster do not always coincide with their exact inner-family resemblance. Within the Germanic cluster, for example, Swedish is more related to Dutch than to Danish, which is counterintuitive. However, within the Slavic cluster languages are grouped correctly (i.e., Slovene-Bulgarian are both South-Slavic) – see [52] for a thorough discussion of GSDN-based language classifications.

In any event, the results of our pilot study show that network-based language classification is a promising approach. At the same time, an F-score of 1 is a high barrier to be mastered by a social-ontology-based approach, which is evaluated next.

6.2. *Testing the Variant of the Sapir-Whorf Hypothesis: Language Classification based on Social Ontologies*

In regards to social ontologies, our version of the SWH contends that language imprints on distributed cognition in such a way that related languages of the same genealogical family are manifested by structurally similar social ontologies (see Section 2). Conversely, our hypothesis implies that unrelated languages, which belong to different genealogical families, result in dissimilar topologies of the corresponding ontologies. We are now in a position to test this

[14] Note that we use the γ-coefficient *only* in the case of those networks that have at least 2,000 vertices as displayed in Table 4.

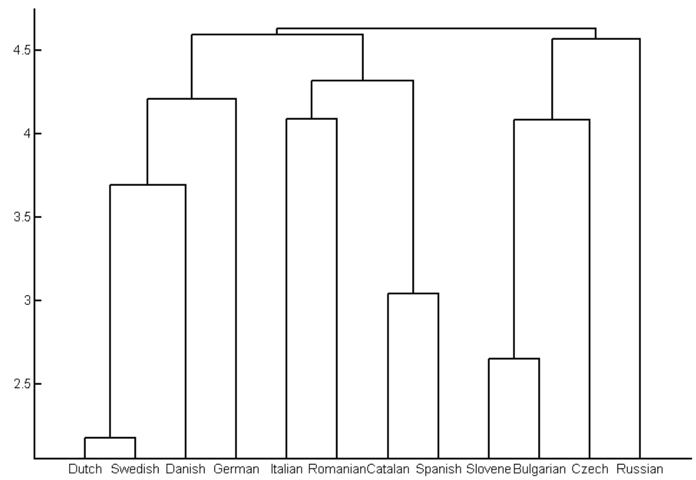


Figure 7: The dendrogram of the best performing classification of 12 languages into three classes in experiment E0 (see Table 5).

Table 5: Experiment E0 testing a version of the Sapir-Whorf Hypothesis: *F*-scores of classifying 12 languages into 3 families based on GSDNs using 24 indices from *Information Theory* and *Network Theory*.

procedure	***F*-score**	**scope**	**source**
`QNA[Mahalanobis,hierarchical,complete]`	1	8/24	IT & NT
`QNA[Correlation,hierarchical,single]`	.63248	24/24	IT & NT
`AVG`	.81624	over non-random approaches	
`random baseline II`	.553	known partition	
`random baseline I`	.54	equi-partition	

hypothesis based on our model of *Generalized Nearly Acyclic Graph*s (GNAG), their quantitative fingerprints and *Quantitative Network Analysis* (QNA) (as described in Section 5.1–5.3).

We start with a reconstruction of the pilot study in Section 6.1. That is, we refer to exactly the same 12 languages as in experiment E0 (see Tables 5 and 2), however, we use GNAGs (as models of social ontologies) instead of GSDNs to obtain a representation model of these languages (see Section 5). This is done to test the expressiveness of GNAGs as input to QNA compared to the more classical approach based on GSDNs. A negative result would mean that the classification based on dependency networks outperforms the one based on social ontologies. The result would even be worse if the latter approach performs as inefficiently as the corresponding baseline scenario. In this case, the similarities of the topologies of social ontologies would tell us nothing about the family resemblances of the corresponding languages.

Table 6 shows that the opposite is true: on the one hand, we obtain the

Table 6: Experiment E1 on a version of the Sapir-Whorf Hypothesis: F-scores of classifying 12 languages into 3 families based on social ontologies by means of 4 classes of topological indices from *Sensitivity Analysis* (SA), *Graph Theory* (GNAG), *Information Theory* (IT) and *Network Theory* (NT).

procedure	F-score	scope	source	class
`QNA[Mahalanobis,hierarchical,complete]`	1.0	7/34	SA	4
`QNA[std. Euclidean,hier.,complete]`	.52381	34/34	SA	4
`QNA[std. Euclidean,hier.,complete]`	.6963	22/52	GNAG	3
`QNA[Euclidean,hierarchical,single]`	.51429	52/52	GNAG	3
`QNA[Euclidean,hierarchical,Ward]`	.67424	17/45	IT	2
`QNA[std. Euclidean,hier.,average]`	.5812	45/45	IT	2
`QNA[std. Euclidean,hier.,average]`	.8381	5/12	NT	1
`QNA[correlation,hierarchical,complete]`	.4963	12/12	NT	1
`QNA[correlation,hierarchical,complete]`	.51852	109/109	all features	
`AVG` (over non-random approaches)	.6492			
`random baseline II`	.553	known partition		
`random baseline I`	.54	equi-partition		

result that the F-score of social-ontology-based classifications is on average (.6492) nearly 10% above the corresponding baselines of .553 and .54. Moreover, all three language families are perfectly separated if a search on the best performing subset of topological indices in Class 4 (*Sensitivity Analysis* – SA) is performed by means of a genetic search algorithm. In this case, we calculate an F-score of 1. This highest possible F-score is computed by means of 7 features only. These are Newman's assortativity index, the graph centrality, the entropy of the standardized closeness centrality, the entropy (variance) of the (cumulative) distribution of geodesic root-related distances, the spheral graph entropy of Bonchev [92], and Dehmer's [74] graph entropy based on linearly decreasing weights. Figure 8 displays the dendrogram, which results from performing experiment E1 by means of these indices: while the group of Romanic languages seems to be plausibly ordered, the Germanic and the Slavic group are not. Interestingly, this dendrogram groups Dutch and Swedish near to each other just as the GSDN-based dendrogram in Figure 7 (although in both cases this is counterintuitive from the point of view of genealogy). In any event, an F-score of 1 is beyond what could be initially expected. As one cannot perform better than by an F-score of 1, this is an argument in support of our approach. Note that if we take all 34 indices of Class 4 into account, the F-score falls to 52% (below both baselines). Once again, there are many features in this set of indices which negatively affect the separation of the focal classes. This observation is recurrent (in all experiments E0-E6) so that sensitivity analyses are an indispensable ingredient of the sort of classification considered here.

Though on a lower level, the same relation (between the full range of indices and its best performing subset) appears in case of Class 1 indices (based on *Network Theory* – NT), Class 2 indices (based on *Information Theory* – IT), and Class 3 indices (based on GNAGs): if we perform a genetic search of the

Table 7: Experiment E2 on a version of the Sapir-Whorf Hypothesis: F-scores of classifying 28 languages into 3 families based on social ontologies by means of 4 classes of topological indices from *Sensitivity Analysis* (SA), *Graph Theory* (GNAG), *Information Theory* (IT) and *Network Theory* (NT).

procedure	F-score	scope	source	class
`QNA[Mahalanobis,hierarchical,complete]`	.78223	26/34	SA	4
`QNA[std. Euclidean,hier.,complete]`	.50022	34/34	SA	4
`QNA[Mahalanobis,hierarchical,complete]`	.72801	18/52	GNAG	3
`QNA[cosine,hierarchical,complete]`	.50866	52/52	GNAG	3
`QNA[Mahalanobis,hierarchical,complete]`	.68052	18/45	IT	2
`QNA[correlation,hierarchical,single]`	.50597	45/45	IT	2
`QNA[correlation,hierarchical,weighted]`	.61267	4/12	NT	1
`QNA[correlation,hierarchical,single]`	.50022	12/12	NT	1
`QNA[correlation,hierarchical,complete]`	.49366	109/109	all features	
`AVG` (over non-random approaches)	.5902			
`random baseline II`	.47214	known partition		
`random baseline I`	.4725	equi-partition		

best performing subset of topological indices, we get an F-score of around 69% in the case of GNAG-related indices and of 67% in the case of IT-related indices. If we do the same in the case of NT-related indices of Class 1, we get a much higher F-score of more than 83%. That is, by exploring only five indices, we classify up to 83% (or ten of twelve languages) correctly. These NT-related features are *not* the usual suspects: once again, this is Newman's assortativity index [1] together with the expected geodesic distance in corresponding regular and random graphs of equal order, the diameter, and the (weighted) cluster coefficient [62].

Obviously, this is a very compact and space efficient representation of structures as complex as social ontologies. Thus, it is a good choice to use this feature model if time and space are critical parameters. However, if one needs to combine space efficiency with classification accuracy, then the 7 SA-related indices of Class 4 are the first choice. Note that if we consider all features in a single experiment without any sensitivity analysis, the F-score is half as high as in case of the best classifier and even falls below the baseline.

To summarize our findings in experiment E1, we do *not* falsify our variant of the SWH, but retain it until any later falsification. In other words, the languages considered in experiment E1 (see Table 6 and 2) are distinguished by the topologies of their corresponding social ontologies such that they are classifiable by QNA into 3 families as predicted by our version of the SWH.

The situation is less obvious, if we enlarge the set of languages to be classified. Table 7 and 8 report continuations of experiment E1 by experiments E2 and E3 (for the target languages see Table 2). In these cases, if we classify 28 languages into 3 families according to the similarities of the topologies of their ontologies: here, the highest F-score falls to 78% and, further, to 69%, if we classify 38 languages as listed in Table 2. In both cases, a genetic search of the best

Table 8: Experiment E3 testing a version of the Sapir-Whorf Hypothesis: F-scores of classifying 38 languages into 3 families based on social ontologies by means of 4 classes of topological indices from *Sensitivity Analysis* (SA), *Graph Theory* (GNAG), *Information Theory* (IT) and *Network Theory* (NT).

procedure	F-score	scope	source	class
QNA[Mahalanobis,hierarchical,complete]	.6969	16/34	SA	4
QNA[std. Euclidean,hier.,single]	.49579	34/34	SA	4
QNA[correlation,hierarchical,complete]	.65439	22/52	GNAG	3
QNA[correlation,hierarchical,single]	.49579	52/52	GNAG	3
QNA[Mahalanobis,hierarchical,complete]	.65038	19/45	IT	2
QNA[correlation,hierarchical,single]	.49421	45/45	IT	2
QNA[Mahalanobis,hierarchical,complete]	.56273	2/12	NT	1
QNA[correlation,hierarchical,single]	.49579	12/12	NT	1
QNA[correlation,hierarchical,single]	.49421	109/109	all features	
AVG (over non-random approaches)	.56			
random baseline II	.44965	known partition		
random baseline I	.4511	equi-partition		

performing subset of SA-related indices guarantees the highest F-scores. Tables 7 and 8 also show that IT- and GNAG-related indices perform above 70% and 65%, respectively, where GNAG-related indices perform better than IT-related indices in experiment E2 and E3, although they are outperformed by SA-related features. From the point of view of linguistic modeling, this supports a network model beyond the classical approach in network theory with its focus on simple graphs. In any event, the baseline scenarios are outperformed in experiment E2 and E3 by all approaches considered here – as well as by their average F-score. Note that GNAG- and IT-related indices are better performing in experiment E2 compared to experiment E1, although the set of languages considered in E1 is a subset of those classified in E2. At first glance, this result is surprising. However, it is explained by the usage of a genetic algorithm to search the best performing subset, which does not necessarily output the optimal subset. Thus, our finding may indicate the existence of better performing subsets in experiment E1 than those we found so far.

From the point of view of experiment E2 and E3, we obtain a positive and a negative result: On the one hand, we still have reasonably large F-scores above the baselines. However, if we compare these findings with experiment E1 (see Table 6), we notice a large loss in F-score due to an enlargement of the set of languages being classified. Thus, our approach is still informative about genealogical resemblances of the languages under consideration, but to a lesser degree than expected according to experiment E1 and the results reported by Table 6. In any event, our findings are still higher than what is expected by chance. Note also that it is reasonable to expect better results if we continue to study more expressive and separable topological indices. Again, the values in Table 7 and 8 do not falsify our variant of the SWH. At this point, we are in a position to examine the social ontology-related variant of Nisbett's Hypothesis.

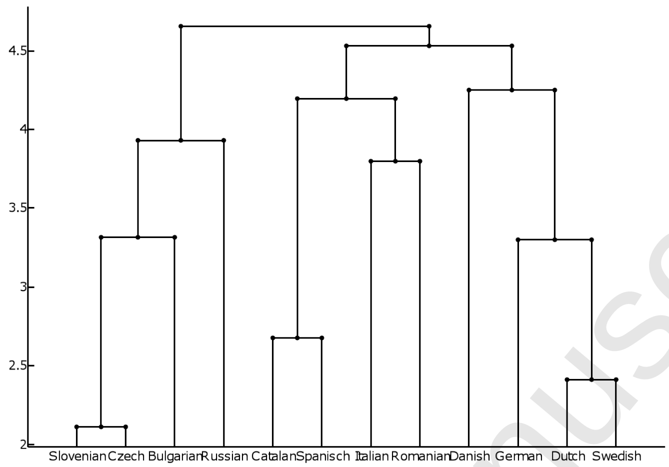


Figure 8: The dendrogram of the best performing classification of 12 languages into three classes in experiment E1 (see Table 6).

6.3. Testing Nisbett's Hypothesis

Our findings regarding the variant of the SWH do not tell us anything about the validity of Nisbett's Hypothesis (see Section 5.3), quite simply as experiments E1–E3 only consider Western languages. However, it is more likely that Nisbett's Hypothesis stands up to falsification, if this also holds for our variant of the SWH. Basically, this expectation is supported by three experiments on Nisbett's Hypothesis as summarized in Tables 9, 10 and 11.

Table 9 starts by separating 3 Sinitic languages and 1 Japonic language from all 38 Western languages that have been considered in experiment E3 (see Table 2). First, we observe a good classification with an F-score of nearly 95%, if we select, once more, a subset of SA-related features by a genetic search.[15] Secondly, we observe an F-score of nearly 90% corresponding to approximately 38 correctly classified languages, if we consider only 6 indices from network theory (feature class 1). As before, this set includes the diameter, the (weighted) cluster coefficient and the expected geodesic distance in corresponding regular and random graphs of equal order, but now supported by the cluster coefficient of [60] and the expected cluster value in a regular graph of equal order.

Furthermore, we see that the baseline scenario that assumes an equi-partition among both target classes is clearly outperformed. However, the random scenario that is informed about the cardinalities of the target classes performs at a high level of nearly 82% – this high random value is due to the largely different sizes of the classes. In any event, experiment E4 does not contradict our variant of Nisbett's Hypothesis. This also holds for Experiment 5 (as summarized in

[15]Note that we consider now a set of 87 indices as elements of Class 4. These are indices, which are degenerated by at most 5% (see Section 5.2).

Table 9: Experiment E4 testing a version of Nisbett's Hypothesis: F-scores of classifying 42 languages into Western and Eastern languages based on their social ontologies by means of 4 classes of topological indices from *Sensitivity Analysis* (SA), *Graph Theory* (GNAG), *Information Theory* (IT) and *Network Theory* (NT). The Eastern class includes 3 Sinitic and 1 Japonic language.

procedure	**F-score**	**scope**	**source**	**class**
`QNA[Mahalanobis,hierarchical,Ward]`	.94505	41/87	SA	4
`QNA[correlation,hierarchical,single]`	.9085	34/87	SA	4
`QNA[cosine,hierarchical,single]`	.9085	26/52	GNAG	3
`QNA[correlation,hierarchical,average]`	.87322	52/52	GNAG	3
`QNA[Euclidean,hierarchical,average]`	.92393	18/45	IT	2
`QNA[correlation,hierarchical,single]`	.86443	45/45	IT	2
`QNA[Mahalanobis,hierarchical,complete]`	.9085	6/12	NT	1
`QNA[correlation,hierarchical,single]`	.9085	12/12	NT	1
`QNA[correlation,hierarchical,single]`	.9085	109/109	all features	
`AVG` (over non-random approaches)	.9055			
`random baseline II`	.81954	known partition		
`random baseline I`	.64368	equi-partition		

Table 10), which additionally considers Korean as an Eastern language – in accordance with Nisbett [28]. We even observe a small gain in F-score, which means that both target classes are better separable if Korean is considered too. The F-scores are much higher than the corresponding random baselines so that we still view our variant of Nisbett's Hypothesis as being *not* falsified.

Next we consider experiment E6 as summarized in Table 11. It continues experiment E5 by additionally viewing 3 Sundic languages as representatives of the group of Eastern languages in the sense of Nisbett. Actually, this extension is excluded by Nisbett, since these Sundic languages have not been influenced in the same ways as the Sinitic, Japonic and Korean languages considered here. Thus, we expect a larger loss in F-score that questions this extension of the class of Eastern languages. This is, in fact, reported by Table 11. In experiment E6, the difference between the best performing classification, on the one hand, and the best performing baseline, on the other, is less then 10%. If we look back at Table 8, we see that in this worst performing experiment on the SWH, the corresponding difference is more than 20% and, thus, much larger. Therefore, we conclude that there is a higher loss in F-score, if we make the questionable extension of the group of Eastern languages (in the sense of Nisbett) by Sundic languages – in accordance to what is predicted by Nisbett's Hypothesis.

All in all, the experiments E4-E6 do not falsify our variant of Nisbett's Hypothesis, and thus we retain it. This means that Western and Eastern languages are distinguishable by topological dissimilarities of their Wikipedia-based social ontologies. This is a new and certainly unexpected result from the point of view of language classification, which – together with the experiments on our variant of the SWH – demonstrates the power of network-theoretical analyses of linguistic systems.

Table 10: Experiment E5 testing a version of Nisbett's Hypothesis, which extends experiment 4 by additionally considering Korean as an Eastern language.

procedure	F-score	scope	source	class
`QNA[Mahalanobis,hierarchical,weighted]`	.94827	42/87	SA	4
`QNA[correlation,hierarchical,single]`	.87829	34/87	SA	4
`QNA[Mahalanobis,hierarchical,single]`	.87829	27/52	GNAG	3
`QNA[correlation,hierarchical,average]`	.84481	52/52	GNAG	3
`QNA[Mahalanobis,hierarchical,Ward]`	.89654	20/45	IT	2
`QNA[correlation,hierarchical,single]`	.84218	45/45	IT	2
`QNA[Mahalanobis,hierarchical,single]`	.87829	5/12	NT	1
`QNA[correlation,hierarchical,single]`	.87829	12/12	NT	1
`QNA[correlation,hierarchical,single]`	.87829	109/109	all features	
`AVG` (over non-random approaches)	.8804			
`random baseline II`	.80422	known partition		
`random baseline I`	.63383	equi-partition		

Table 11: Experiment E6 testing a version of Nisbett's Hypothesis, which extends experiment 5 by additionally considering 3 Sundic languages as Eastern languages.

procedure	F-score	scope	source	class
`QNA[Euclidean,hierarchical,average]`	.85109	39/87	SA	4
`QNA[correlation,hierarchical,single]`	.80236	34/87	SA	4
`QNA[Mahalanobis,hierarchical,Ward]`	.81794	26/52	GNAG	3
`QNA[correlation,hierarchical,single]`	.78901	52/52	GNAG	3
`QNA[Euclidean,hierarchical,average]`	.85109	20/45	IT	2
`QNA[correlation,hierarchical,single]`	.78901	45/45	IT	2
`QNA[correlation,hierarchical,single]`	.80236	6/12	NT	1
`QNA[correlation,hierarchical,single]`	.80236	12/12	NT	1
`QNA[correlation,hierarchical,single]`	.80236	109/109	all features	
`AVG` (over non-random approaches)	.8120			
`random baseline II`	.75304	known partition		
`random baseline I`	.62477	equi-partition		

6.4. Discussion

Before we start a more general discussion of our findings, we hint at two characteristics of our numerical results. Firstly, if we compare the feature classes 1, 2 and 3 and disregard Class 4 of SA-related features for a while, we see that GNAG-related features mostly perform best in our experiments on the SWH, while IT-related features perform better in our experiments on Nisbett's Hypothesis. At least from the point of view of experiment E2-E3 this means that social ontologies are better separated by means of indices, which reflect their characteristics in terms of generalized acyclic graphs. This is an argument in favor of more informative graph models beyond the simple graphs traditionally analyzed in complex network theory [71].

Secondly, our results show that selections of indices according to Konstantinova's index of degeneration (see Section 5.2) perform best if being combined with a sensitivity analysis. This selection is deterministic as it selects all indices with a sensitivity of at least 95% or even of 100% as in the case of experiments on the SWH. That is, only indices, which in a reference corpus of 160 ontologies separate at least 152 graphs correctly, are collected in Class 4 of SA-related features. Because of this determinism, the selection can be automatized. This is a strong argument to look for more expressive sensitivity analyses, which may help to improve network-based structural classification.

Generally speaking, our reasons to apply network theory in the area of language classification can be summarized as follows:

1. Firstly, our aim is to model linguistic structures beyond tree-like graphs. We aim to explore systems, which recently evolved in web-based communication. These systems are characterized by the networking of hundreds and thousands of vertices beyond tree-like models to which linguistics traditionally pertains. In this sense, the networking of web-based units relates to a rapidly emerging field of linguistic manifestation. The present article has shown that this networking is even indicative of family resemblances of languages. So network models of the sort presented here are interesting for general linguistics – at least as comparative studies.
2. Secondly, we stress the expressiveness of structural models in classifying linguistic units beyond content-based models traditionally used in computational linguistics [93]. This accentuation of structure modeling is in line with Dimter's [25] experiment on text typology and its algorithmic reconstruction [26]. Dimter shows that, obviously, structure is an underestimated source of identifying linguistics types. We extend this approach to linguistic networks and show that purely structure-based classifications are successful in this area too. This raises the question about the expressiveness of structure-based classifications in computational linguistics in general to which our article contributes.
3. Thirdly, our experiment complements recent approaches to use web-based resources in NLP. These approaches have in common that they explore the structure of Wikipedia and related resources to derive representation models in text categorization [94], to compute semantic relatedness [95], or to induce topic labels [96]. Based on our findings, we get a first insight into the context-sensitivity of such approaches. That is, we observe that networks of different language families vary to an extent that makes them automatically separable. If this finding is continuously confirmed, algorithms for NLP, which structurally explore such resources, become context-dependent – at least on the level of the underlying language family. In such a case, the average geodesic distance, for example, would mean something else, say, in Sundic vs. Slavic linguistic networks. Following this line of research, network-theoretical research as the one presented here can contribute to NLP.

Generally speaking, our findings indicate the reliability of a novel source of language classification based on human computation as manifested by wiki media. Other than the (e.g., graphemical, morphological, lexical, or syntactic) representation models traditionally used for genealogical classification, we successfully classify languages by means of resources of the social web. One might object that our approach is lexical as it starts with exploring conceptual systems manifested by lexemes. However, this is not true as we only explore structural characteristics of these resources, while we disregard any content units. To the best of our knowledge this is the first such approach to language classification.

7. Conclusion

In this article, we presented a network-theoretical approach to language classification. Our study is a first attempt to classify languages by means of the topological characteristics of social ontologies generated in these languages. We have tested two related hypotheses: a variant of the Sapir-Whorf Hypothesis and a variant of Nisbett's Hypothesis on differences in Western and Eastern cultures. In this way, we gained access to structural analyses of linguistic networks by example of Wikipedia-based social ontologies as a new resource of language classification.

In support of the SWH, we successfully classified languages into three genealogical groups. We also outperformed corresponding baselines of random classification. Concerning Nisbett's variant of the SWH, we obtained a similar result by separating Western and Eastern languages. As predicted by Nisbett, the classification worsened by extending the corpus of Eastern languages by Sundic languages. In any event, enlarging the number of classes may worsen our results as well as we observed in our experiments. Obviously, the results obtained could have been biased by the number of classes and related factors such as the size of the language families, the validity of the underlying corpora and the independence of the data sources. Thus, we aim to examine these factors in further studies to undermine our findings. Additionally, future work will address the construction of more elaborate baselines, and checking the extensibility of our approach to other kinds of social ontologies. Further, we plan to build more expressive graph models in conjunction with topological indices that are more separable to get better classification results. We also want to extend sensitivity analyses as the one based on Konstantinova's index of degeneration to get classifiers that can be reliably transferred to other areas of linguistic networks. Finally, we will make larger classification experiments to extend the range of language families covered by our approach.

Acknowledgement

Financial support of the German Federal Ministry of Education and Research (BMBF) through the project *Linguistic Networks* (`www.linguistic-networks.net`), of the German Research Foundation (DFG) through the Cluster of Excellence *Cognitive Interaction Technology* and the Collaborative Research Centre *Alignment in Communication* is gratefully acknowledged. We also thank Daniel

Kinzler of the German Wikimedia Foundation for fruitful hints on the organization of the category system of Wikipedia.

References

[1] M. E. J. Newman, The structure and function of complex networks, SIAM Review 45 (2003) 167–256.

[2] R. Ferrer i Cancho, R. V. Solé, R. Köhler, Patterns in Syntactic Dependency-Networks, Physical Review E 69 (5) (2004) 051915.

[3] A. Mehler, Large Text Networks as an Object of Corpus Linguistic Studies, in: A. Lüdeling, M. Kytö (Eds.), Corpus Linguistics. An International Handbook of the Science of Language and Society, De Gruyter, Berlin/New York, 328–382, 2008.

[4] B. Whorf, Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf, MIT Press, Cambridge, 1956.

[5] J. A. Lucy, Language Diversity and Thought. A reformulation of the linguistic relativity hypothesis, Cambridge University Press, Cambridge, 1992.

[6] J. F. Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks/Cole, Pacific Grove, 2000.

[7] B. Leuf, W. Cunningham, The Wiki Way. Quick Collaboration on the Web, Addison Wesley, 2001.

[8] P. Mika, A. Gangemi, Descriptions of Social Relations, in: Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web, 2004.

[9] J. R. Searle, Social Ontology. Some Basic Principles, Anthropological Theory 6 (1) (2006) 12–29.

[10] M. H. Bickhard, Social Ontology as Convention, Topoi 27 (1-2) (2008) 139–149.

[11] L. Steels, Collaborative tagging as distributed cognition, Pragmatics & Cognition 14 (2) (2006) 287–292.

[12] J. Voss, Collaborative thesaurus tagging the Wikipedia way, arXiv.org:cs/0604036, 2006.

[13] J. Hollan, E. Hutchins, D. Kirsh, Distributed cognition: toward a new foundation for human-computer interaction research, ACM Transaction on Computer-Human Interaction 7 (2) (2000) 174–196.

[14] J. A. Lucy, Linguistic Relativity, Annual Review of Anthropology 26 (1997) 291–312.

[15] S. Pinker, The Language Instinct: How the Mind Creates Language, Perennial, 1994.

[16] P. Lee, The Whorf Theory Complex – A Critical Reconstruction, John Benjamins, 1996.

[17] E. H. Lenneberg, J. M. Roberts, The language of experience: A study in methodology, International Journal of American Linguistics Supplement to Volume 22.

[18] C. L. Hardin, L. Maffi (Eds.), Color Categories in Thought and Language, Cambridge University Press, 1997.

[19] T. Regier, P. Kay, N. Khetarpal, Color naming reflects optimal partitions of color space, Proceedings of the National Academy of Sciences 104 (2007) 1436–1441.

[20] M. Bowermann, The Origins of children's spatial semantic categories: cognitive versus linguistic determinants, in: J. J. Gumperz, S. C. Levinson (Eds.), Rethinking linguistic relativity, Cambridge University Press, 145–176, 1996.

[21] S. C. Levinson, Frames of reference and Molyneux's question: Cross-linguistic evidence, in: P. Bloom, M. Peterson, L. Nadel, M. Garrett (Eds.), Language and space, MIT press, Cambridge, 109–169, 1996.

[22] J. A. Lucy, S. Gaskins, Grammatical categories and the development of classification preferences: a comparative approach, in: M. Bowerman, S. Levinson (Eds.), Language Acquisition and Conceptual Development, Cambridge University Press, 257–283, 2001.

[23] L. Boroditsky, Does Language Shape Thought? Mandarin and English Speakers' Conceptions of Time, Cognitive Psychology 43 (2001) 1–22.

[24] H. Liiv, J. Tuldava, On Classifying Texts with the Help of Cluster Analysis, in: L. Hřebíček, G. Altmann (Eds.), Quantitative Text Analysis, Wissenschaftlicher Verlag, Trier, 253–262, 1993.

[25] M. Dimter, Textklassenkonzepte heutiger Alltagssprache, Niemeyer, Tübingen, 1981.

[26] A. Mehler, P. Geibel, O. Pustylnikov, Structural Classifiers of Text Types: Towards a Novel Model of Text Representation, Journal for Language Technology and Computational Linguistics (JLCL) 22 (2) (2007) 51–66.

[27] O. Pustylnikov, A. Mehler, Structural Differentiae of Text Types. A Quantitative Model, in: Proceedings of the 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKl), 655–662, 2007.

[28] R. E. Nisbett, The Geography of Thought. How Asians and Westeners Think Differently ... and Why, Free Press, New York, 2003.

[29] L. McDonough, S. Choi, J. Mandler, Development of language-specific categorization of spatial relations from pre-linguistic to linguistic stage a preliminary study, Presented at the Finding the Words Conference at Stanford University, 2000.

[30] D. Casasanto, Who's Afraid of the Big Bad Whorf? Crosslinguistic Differences in Temporal Language and Thought, Language Learning 58 (1) (2008) 63–79.

[31] J. Hurford, Nativist and functional explanations in language acquisition, in: I. M. Roca (Ed.), Logical Issues in Language Acquisition, Foris, Dordrecht, 85–136, 1990.

[32] M. Swadesh, Lexico-statistic dating of prehistoric ethnic contacts, in: Proceedings of the American philosophical society, vol. 96, 452–463, 1952.

[33] S. C. Gudschinsky, The ABC's of Lexicostatistics (Glottochronology), Word 12 (2) (1956) 175–210.

[34] R. L. Oswalt, The Detection of Remote Linguistic Relationships, Studies in the Humanities and Verbal Behavior 3 (1970) 117–129.

[35] M. A. Covington, An algorithm to align words for historical comparison, Computational Linguistics 22 (4) (1996) 481–496.

[36] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Doklady Akademii Nauk SSSR 163 (4) (1965) 845–848, english in: Soviet Physics Doklady, 10 (8) (1966) 707–710.

[37] R. A. Wagner, M. J. Fischer, The String-to-String Correction Problem, Journal of the Association for Computing Machinery 21 (1) (1974) 168–173.

[38] G. Kondrak, Algorithms for Language Reconstruction, Ph.D. thesis, University of Toronto, 2002.

[39] B. Kessler, Phonetic Comparison Algorithms, Transactions of the Philological Society 103 (2) (2005) 243–260.

[40] E. Sapir, Time Perspective in Aboriginal American Culture, A Study in Method, Geological Survey of Canada, Memoir 90, Anthropological Series No. 13, Canada, Department of Mines, Ottawa, 1916.

[41] K. Bergsland, H. Vogt, On the Validity of Glottochronology, Current Anthropology 3 (2) (1962) 115–153.

[42] S. M. Embleton, Statistics in Historical Linguistics, vol. 30 of *Quantitative Linguistics*, Studienverlag Dr. N. Brockmeyer, Bochum, 1986.

[43] M. Pagel, Q. D. Atkinson, A. Meade, Frequency of word-use predicts rates of lexical evolution throughout Indo-European history, Nature 449 (7163) (2007) 717–720.

[44] T. M. Ellison, S. Kirby, Measuring Language Divergence by Intra-Lexical Comparison, in: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, ACL, 273–280, 2006.

[45] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang, M. A. Nowak, Quantifying the evolutionary dynamics of language, Nature 449 (7163) (2007) 713–716.

[46] E. W. Holman, S. Wichmann, C. H. Brown, V. Velupillai, A. Müller, D. Bakker, Explorations in automated language classification, Folia Linguistica 42 (2) (2008) 331–354.

[47] T. Warnow, S. N. Evans, D. Ringe, L. Nakhleh, A Stochastic model of language evolution that incorporates homoplasy and borrowing, in: Phylogenetic Methods and the Prehistory of Languages, chap. 7, 75–87, 2006.

[48] R. F. Port, A. P. Leary, Against Formal Phonology, Language 81 (4) (2005) 927–964.

[49] W. B. Cavnar, J. M. Trenkle, N-Gram-Based Text Categorization, in: In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 161–175, 1994.

[50] H. Daumé III, Non-Parametric Bayesian Areal Linguistics, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Boulder, Colorado, 593–601, 2009.

[51] A. Mukherjee, M. Choudhury, A. Basu, N. Ganguly, Emergence of community structures in vowel inventories: an analysis based on complex networks, in: Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, Prague, 2007.

[52] O. Pustylnikov, A. Mehler, Typology by means of Language Networks. Enhancing Typological Methods by an Integrated View on Language, in preparation, 2010.

[53] U. Brandes, M. Eiglsperger, I. Herman, M. Himsolt, M. S. Marshall, GraphML Progress Report: Structural Layer Proposal, in: Proc. 9th Intl. Symp. Graph Drawing (GD '01), Lecture Notes in Computer Science 2265, Springer, 501–512, 2002.

[54] F. Harary, Graph Theory, Addison Wesley, Boston, 1969.

[55] A. Mehler, Generalized Shortest Paths Trees: A Novel Graph Class Applied to Semiotic Networks, in: M. Dehmer, F. Emmert-Streib (Eds.), Analysis of Complex Networks: From Biology to Linguistics, Wiley-VCH, Weinheim, 175–220, 2009.

[56] J. Bang-Jensen, G. Gutin, Digraphs. Theory, Algorithms and Applications, Springer, London/Berlin, 2006.

[57] A. Mehler, Structural Similarities of Complex Networks: A Computational Model by Example of Wiki Graphs, Applied Artificial Intelligence 22 (7&8) (2008) 619–683.

[58] M. Dehmer, A. Mehler, F. Emmert-Streib, Graph-theoretical Characterizations of Generalized Trees, in: Proceedings of the 2007 International Conference on Machine Learning: Models, Technologies & Applications (MLMTA'07), June 25-28, 2007, Las Vegas, 113–117, 2007.

[59] A. Mehler, Structure Formation in the Web. A Graph-Theoretical Model of Hypertext Types, in: A. Witt, D. Metzing (Eds.), Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology, Text, Speech and Language Technology, Springer, Dordrecht, 2009.

[60] D. J. Watts, S. H. Strogatz, Collective Dynamics of 'Small-World' Networks, Nature 393 (1998) 440–442.

[61] B. Bollobás, O. M. Riordan, Mathematical Results on Scale-Free Random Graphs, in: S. Bornholdt, H. G. Schuster (Eds.), Handbook of Graphs and Networks. From the Genome to the Internet, Wiley-VCH, Weinheim, 1–34, 2003.

[62] M. Á. Serrano, M. Boguñá, R. Pastor-Satorras, Correlations in weighted networks, Physical Review E 74 (2006) 055101.

[63] A.-L. Barabási, R. Albert, Emergence of Scaling in Random Networks, Science 286 (1999) 509–512.

[64] M. Dehmer, A. Mowshowitz, A Natural History of Graph Entropy, submitted, 2009.

[65] J. J. Freyd, Shareability: the social psychology of epistemology, Cognitive Science 7 (1983) 191–210.

[66] R. A. Botafogo, E. Rivlin, B. Shneiderman, Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics, ACM Transactions on Information Systems 10 (2) (1992) 142–180.

[67] R. Feldman, J. Sanger, The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, Cambridge, 2007.

[68] A. Mehler, A Quantitative Graph Model of Social Ontologies by Example of Wikipedia, in: M. Dehmer, F. Emmert-Streib, A. Mehler (Eds.), Towards an Information Theory of Complex Networks: Statistical Methods and Applications, Birkhäuser, Boston/Basel, 2010.

[69] E. V. Konstantinova, V. A. Skorobogatov, M. V. Vidyuk, Applications of information theory in chemical graph theory, Indian journal of chemistry. Sect. A: Inorganic, physical, theoretical & analytical 42 (6) (2003) 1227–1240.

[70] M. Dehmer, K. Varmuza, S. Borgert, F. Emmert-Streib, On entropy-based molecular descriptors: statistical analysis of real and synthetic chemical structures, Journal of chemical information and modeling 49 (7) (2009) 1655–1663.

[71] G. Caldarelli, A. Vespignani (Eds.), Large Scale Structure and Dynamics of Complex Networks, World Scientific, New Jersey, 2007.

[72] A. Barrat, M. Barthélemy, A. Vespignani, Dynamical Processes on Complex Networks, Cambridge University Press, Cambridge, 2008.

[73] S. Wasserman, K. Faust, Social Network Analysis. Methods and Applications, Cambridge University Press, Cambridge, 1999.

[74] M. Dehmer, Information Processing in Complex Networks: Graph Entropy and Information Functionals, Applied Mathematics and Computation 201 (2008) 82–94.

[75] P. Gärdenfors, Conceptual Spaces, MIT Press, Cambridge, MA, 2000.

[76] T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley-Interscience, Hoboken, 2006.

[77] G. Altmann, W. Lehfeldt, Allgemeine Sprachtypologie, Fink, München, 1973.

[78] R. Ferrer i Cancho, R. V. Solé, R. Köhler, Patterns in syntactic dependency networks, Physical Review E 69 (2004) 051915.

[79] S. Wallis, Searching treebanks and other structured corpora, in: A. Lüdeling, M. Kytö (Eds.), Corpus Linguistics: An International Handbook, De Gruyter, Berlin/New York, 2008.

[80] L. van der Beek, G. Bouma, R. Malouf, G. van Noord, The Alpino dependency treebank, in: Proc. of the Conf. on Computational Linguistics in the Netherlands (CLIN '02), 2002.

[81] M. T. Kromann, The Danish Dependency Treebank and the underlying linguistic theory, in: J. Nivre, E. Hinrichs (Eds.), Proc. of TLT 2003, Växjö University Press, 2003.

[82] F. Hristea, M. Popescu, A dependency grammar approach to syntactic analysis with special reference to Romanian, in: Building Awareness in Language Technology, University of Bucharest Publishing House, 2003.

[83] I. Boguslavsky, I. Chardin, S. Grigorieva, N. Grigoriev, L. Iomdin, L. Kreidlin, N. Frid, Development of a Dependency Treebank for Russian and its Possible Applications in NLP, in: Proceedings of the 3rd International Conference on Language Ressources and Evaluation (LREC 2002), Las Palmas, Gran Canaria, 2002.

[84] S. Džeroski, T. Erjavec, N. Ledinek, P. Pajas, Z. Žabokrtský, A. Žele, Towards a Slovene dependency treebank, in: Proc. of LREC 2006, 2006.

[85] J. Nivre, J. Nilsson, J. Hall, Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation, in: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006), May 24-26, Genua, Italy, 2006.

[86] C. Bosco, V. Lombardo, D. Vassallo, L. Lesmo, Building a treebank for Italian: a data-driven annotation schema, in: Proc. of LREC 2000, 2000.

[87] M. Civit, N. Bufí, P. Valverde, Cat3LB: a Treebank for Catalan with Word Sense Annotation, in: TLT2004, Tübingen University, 27–38, 2004.

[88] M. Civit, M. Martí, Building Cast3LB: A Spanish Treebank, a Research on Language and Computation, Springer Verlag (2005) 549–574.

[89] J. Hajič, Building a Syntactically Annotated Corpus: The PragueDependency Treebank, in: E. Hajičová (Ed.), Issues of Valency and Meaning. Studies in Honour of Jarmila-Panevová, Karolinum, Charles University Press, Prague, Czech Republic, 106–132, 1998.

[90] P. Osenova, K. Simov, BTB-TR05: BulTreeBank Stylebook. BulTreeBank Project Technical Report Nr. 05, Tech. Rep., Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, 2004.

[91] S. Brants, S. Dipper, S. Hansen, W. Lezius, G. Smith, The TIGER Treebank, in: Proceedings of the Workshop on Treebanks and Linguistic Theories, Sozopol, 2002.

[92] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, Research Studies Press, Chichester, 1983.

[93] C. Kemp, J. B. Tenenbaum, The discovery of structural form, Proceedings of the National Academy of Sciences 105 (31) (2008) 10687–10692.

[94] E. Gabrilovich, S. Markovitch, Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge, in: Proceedings of the Twenty-First National Conference on Artificial Intelligence, Boston, MA, 2006.

[95] E. Yeh, D. Ramage, C. D. Manning, E. Agirre, A. Soroa, WikiWalk: Random walks on Wikipedia for Semantic Relatedness, in: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4), Association for Computational Linguistics, Suntec, Singapore, 41–49, 2009.

[96] U. Waltinger, A. Mehler, Social Semantics and its Evaluation by Means of Semantic Relatedness and Open Topic Models, in: IEEE/WIC/ACM International Conference on Web Intelligence, September 15–18, Milano, 2009.