Sabine Schulte im Walde, Stefan Müller

Using Web Corpora for the Automatic Acquisition of Lexical-Semantic Knowledge

This article presents two case studies to explore whether and how web corpora can be used to automatically acquire lexical-semantic knowledge from distributional information. For this purpose, we compare three German web corpora and a traditional newspaper corpus on modelling two types of semantic relatedness: (1) Assuming that free word associations are semantically related to their stimuli, we explore to which extent stimulus–associate pairs from various associations norms are available in the corpus data. (2) Assuming that the distributional similarity between a noun–noun compound and its nominal constituents corresponds to the compound's degree of compositionality, we rely on simple corpus co-occurrence features to predict compositionality. The case studies demonstrate that the corpora can indeed be used to model semantic relatedness, (1) covering up to 73/77% of verb/noun–association types within a 5-word window of the corpora, and (2) predicting compositionality with a correlation of $\rho = 0.65$ against human ratings. Furthermore, our studies illustrate that the corpus parameters *domain, size* and *cleanness* all have an effect on the semantic tasks.

1 Motivation

Distributional models assume that the contexts of a linguistic unit (such as a word, a multi-word expression, a phrase, a sentence, etc.) provide information about the meaning of the linguistic unit (Firth, 1957; Harris, 1968). They have been widely applied in data-intensive lexical semantics (among other areas), and proven successful in diverse research issues, such as the representation and disambiguation of word senses (Schütze, 1998; McCarthy et al., 2004; Springorum et al., 2013), selectional preference modelling (Herdagdelen and Baroni, 2009; Erk et al., 2010; Schulte im Walde, 2010), compositionality of compounds and phrases (McCarthy et al., 2003; Reddy et al., 2011; Boleda et al., 2013; Schulte im Walde et al., 2013), or as a general framework across semantic tasks (Baroni and Lenci, 2010; Padó and Utt, 2012), to name just a few examples.

While distributional models are well-established in computational linguistics, from a cognitive point of view the relationship between meaning and distributional models has been more controversial (Marconi, 1997; Glenberg and Mehta, 2008), because distributional models are expected to cover the linguistic ability of how to use language (*inferential abilities*), but they do not incorporate a knowledge of the world (*referential abilities*). Since distributional models are very attractive – the underlying parameters

being accessible from even low-level annotated corpus data – we are interested in maximising the benefit of distributional information for lexical semantics, and in exploring the potential and the limits with regard to individual semantic tasks. More specifically, our work addresses distributional approaches with respect to semantic relatedness. As resources for the distributional knowledge we rely on web corpora, because (a) these corpora are the largest language corpora currently available, and size matters (Banko and Brill, 2001; Curran and Moens, 2002; Pantel et al., 2004; Hickl et al., 2006), and (b) web corpora are domain-independent, in comparison to e.g. large newspaper collections, and thus cover the potentially widest breadth of vocabulary.

In this article, we present two case studies to explore whether and how web corpora can be used to automatically acquire lexical-semantic knowledge from distributional information. For this purpose, we compare three German web corpora (differing in domain, size and cleanness) on modelling two types of semantic relatedness:

(1) semantic associations, and

(2) the compositionality of noun-noun compounds.

Concerning task (1), we in addition compare the usage of the web corpora against applying a traditional newspaper corpus.

Semantic Relatedness Task (1) assumes that free word associations (i.e., words that are spontaneously called to mind by a stimulus word) represent a valuable resource for cognitive and computational linguistics research on semantic relatedness. This assumption is based on a long-standing hypothesis, that *associations reflect meaning components of words* (Nelson et al., 1997, 2000; McNamara, 2005), and are thus semantically related to the stimuli. Our first semantic relatedness task will explore to which extent lexical-semantic knowledge – as represented by various collections of associations – is available in the corpus data.

Semantic Relatedness Task (2) is the prediction of compositionality for a set of German noun-noun compounds, i.e., the degree of semantic relatedness between a compound and its constituents. We rely on simple corpus co-occurrence features to instantiate a distributional model of the compound nouns and their nominal constituents, and use the cosine measure to rate the distributional similarity between the compounds and the constituents. Assuming that the distributional similarity between a compound and a constituent corresponds to the compound–constituent degree of compositionality, we compare our predictions against human compositionality ratings.

The remainder of the article is organised as follow. Section 2 first introduces the (web) corpora that are used for our semantic tasks. Section 3 then describes the usage of these corpora with regard to our two tasks, (1) the availability of association knowledge and (2) predicting the compositionality of noun compounds. Section 4 summarises and discusses the results and insights.

2 Corpora

The main goal of our work is to explore the potential and the limits of distributional information for modelling semantic relatedness. We are thus interested in (a) exploiting the largest available corpus data but at the same time (b) comparing the usage of different types of corpora. Accordingly, corpus criteria that are central to our work are

- corpus domain(s) and
- corpus size.

Our case studies in Section 3 make use of the following four corpora:

1. *WebKo*

   The *WebKo* corpus is a slightly cleaned version of *deWaC*, a standard cross-domain German web corpus created by the *WaCky* group (Baroni et al., 2009). The cleaning was performed through simple heuristics to identify and disregard implausible domains and to repair dates and incorrectly separated tokens.

2. *SdeWaC*

   The *SdeWaC* corpus (Faaß et al., 2010; Faaß and Eckart, 2013) is a more severely cleaned version of *WebKo*. The corpus cleaning had focused on removing duplicates from the *deWaC*, and on disregarding sentences that were syntactically ill-formed (following Quasthoff et al. (2006) regarding heuristics such as the number of commas per sentence, the number of spaces in proportion to sentence length, etc.; and relying on a parsability index provided by the standard dependency parser by Schiehlen (2003)). The SdeWaC is freely available and can be downloaded from `http://wacky.sslmit.unibo.it/`.

   A special feature of the SdeWaC is that the sentences in the corpus have been sorted alphabetically, so going beyond the sentence border is likely to entering a sentence that did not originally precede or follow the sentence of interest. This will have an effect on window co-occurrence in Section 3.

3. *German Wikipedia (Wiki-de)*

   The *Wiki-de* corpus is also a German web corpus, based on the official Wikipedia dump[1] `dewiki-20110410` from April 10, 2011. The dump has been downloaded and processed by André Blessing, *Institut für Maschinelle Sprachverarbeitung (IMS)*. The processing was performed using the *Java Wikipedia Library (JWPL)*[2], an open-source Java-based application programming interface to access and parse the information in the Wikipedia articles (Zesch et al., 2008). The encyclopaedic knowledge is expected to complement the knowledge induced from the other two web corpora, cf. Roth and Schulte im Walde (2008).

[1] `http://dumps.wikimedia.org/dewiki/`
[2] `http://www.ukp.tu-darmstadt.de/software/jwpl/`

4. *Huge German Corpus (HGC)*

   The *HGC* corpus is a large collection of German newspaper corpora, containing data from *Frankfurter Rundschau*, *Stuttgarter Zeitung*, *VDI-Nachrichten*, *die tageszeitung (taz)*, *Gesetzestexte (German Law Corpus)*, *Donaukurier*, and *Computerzeitung* from the 1990s. We used the HGC in contrast to the above three web corpora, to explore the influence of the more restricted corpus domain.

Table 1 provides an overview of the corpus sizes. All of the corpora have been tokenised and part-of-speech tagged with the *Tree Tagger* (Schmid, 1994).

	WebKo	SdeWaC	Wiki-de	HGC
sentences	71,585,693	45,400,446	23,205,536	9,255,630
words (tokens)	1,520,405,616	884,356,312	432,131,454	204,813,118
words (types)	14,908,301	9,220,665	7,792,043	3,193,939

Table 1: Overview of corpora.

3 Web Corpora and Semantic Relatedness: Two Case Studies

This section as the main part of the article explores whether and how web corpora can be used to automatically acquire lexical-semantic knowledge from distributional information. For this purpose, we compare the German corpora introduced in Section 2 on modelling two types of semantic relatedness, the availability of semantic associates (Section 3.1) and the prediction of compositionality for German noun-noun compounds (Section 3.2).

3.1 Availability of Semantic Associates in Web Corpora

Our first semantic relatedness task will explore to which extent lexical-semantic knowledge – as represented by three collections of association norms – is available in the (web) corpus data. Section 3.1.1 first introduces the relevant background on association norms, before Section 3.1.2 provides an explicit description of our hypotheses. Section 3.1.3 then describes the actual distributional explorations.

3.1.1 Background and Earlier Work on Association Norms

Associations are commonly obtained by presenting *target stimuli* to the participants in an experiment, who then provide *associate responses*, i.e., words that are spontaneously called to mind by the stimulus words. The quantification of the resulting stimulus–association pairs (i.e., how often a certain association is provided for a certain stimulus) is called *association norm*. Table 2 (as taken from Schulte im Walde et al. (2008)) provides the 10 most frequent associate responses for the ambiguous verb *klagen* and the ambiguous noun *Schloss* as examples.

klagen 'complain, moan, sue'		
Gericht	'court'	19
jammern	'moan'	18
weinen	'cry'	13
Anwalt	'lawyer'	11
Richter	'judge'	9
Klage	'complaint'	7
Leid	'suffering'	6
Trauer	'mourning'	6
Klagemauer	'Wailing Wall'	5
laut	'noisy'	5

Schloss 'castle, lock'		
Schlüssel	'key'	51
Tür	'door'	15
Prinzessin	'princess'	8
Burg	'castle'	8
sicher	'safe'	7
Fahrrad	'bike'	7
schließen	'close'	7
Keller	'cellar'	7
König	'king'	7
Turm	'tower'	6

Table 2: Associate responses and associate frequencies for example stimuli.

Association norms have a long tradition in psycholinguistic research, where the implicit notion that associations reflect meaning components of words has been used for more than 30 years to investigate semantic memory. One of the first collections of word association norms was done by Palermo and Jenkins (1964), comprising associations for 200 English words. The *Edinburgh Association Thesaurus* (Kiss et al., 1973) was a first attempt to collect association norms on a larger scale, and also to create a network of stimuli and associates, starting from a small set of stimuli derived from the Palermo and Jenkins norms. A similar motivation underlies the association norms from the University of South Florida (Nelson et al., 1998),[3] who grew a stimulus-associate network over more than 20 years, from 1973. More than 6,000 participants produced nearly three-quarters of a million responses to 5,019 stimulus words. In another long-term project, Simon de Deyne and Gert Storms are collecting associations to Dutch words, cf. `www.smallworldofwords.com`. Previously, they performed a three-year collection of associations to 1,424 Dutch words (de Deyne and Storms, 2008b). Smaller sets of association norms have also been collected for example for German (Russell and Meseck, 1959; Russell, 1970), Dutch (Lauteslager et al., 1986), French (Ferrand and Alario, 1998) and Spanish (Fernández et al., 2004) as well as for different populations of speakers, such as adults vs. children (Hirsh and Tree, 2001).

In parallel to the interest in collecting association norms, researchers have analysed association data in order to get insight into semantic memory and – more specifically – issues concerning semantic relatedness. For example, Clark (1971) classified stimulus–association relations into sub-categories of paradigmatic and syntagmatic relations, such as synonymy and antonymy, selectional preferences, etc. Heringer (1986) concentrated on syntagmatic associations to a small selection of 20 German verbs. He asked his subjects to provide question words as associations (e.g., *wer* 'who', *warum* 'why'), and used the responses to investigate the valency behaviour of the verbs. Spence and Owens (1990) showed that associative strength and word co-occurrence are correlated. Their

[3] `http://www.usf.edu/FreeAssociation/`

investigation was based on 47 pairs of semantically related concrete nouns, as taken from the Palermo and Jenkins norms, and their co-occurrence counts in a window of 250 characters in the 1-million-word Brown corpus. Church and Hanks (1989) were the first to apply information-theoretic measures to corpus data in order to predict word association norms for lexicographic purposes. Our own work analysed German noun and verb associations at the syntax-semantics interface (Schulte im Walde et al., 2008). Schulte im Walde and Melinger (2008) performed a more in-depth analysis of window co-occurrence distributions of stimulus–response pairs. Roth and Schulte im Walde (2008) explored whether dictionary and encyclopaedic information provides more world knowledge about associations than corpus co-occurrence, and found that the information in the three resource types complements each other.

In experimental psychology, association norms have been used extensively to conduct studies with variations of the semantic priming technique to investigate (among other things) word recognition, knowledge representation and semantic processes (see McNamara (2005) for a review of methods, issues, and findings). In the last decade, association norms have also found their way into lexical-semantic research in computational linguistics. For example, Rapp (2002) developed corpus-based approaches to predict paradigmatic and syntagmatic associations; de Deyne and Storms (2008a) created semantic networks from Dutch associations; and Schulte im Walde (2008) used associations to German verbs to select features for automatic semantic classification.

3.1.2 Associations, Semantic Relatedness, and Corpus Co-Occurrence

Our analyses to follow rely on three well-known hypotheses: (i) the psycholinguistic notion that *associations reflect meaning components of words* (Nelson et al., 1997, 2000; McNamara, 2005); (ii) the *co-occurrence hypothesis* that associations are related to the textual co-occurrence of the stimulus–association pairs (Miller, 1969; Spence and Owens, 1990; McKoon and Ratcliff, 1992; Plaut, 1995); and (iii) the *distributional hypothesis* that contexts of a word provide information about its meaning (Firth, 1957; Harris, 1968). Figure 1 illustrates the triangle that combines the three hypotheses and at the same time bridges the gap between long-standing assumptions in psycholinguistic and computational linguistics research, regarding associations, semantic memory, and corpus co-occurrence.

According to these three hypotheses, association norms thus represent a valuable resource for cognitive and computational linguistics research on semantic relatedness. In the current study, we exploit three collections of association norms to explore the availability of lexical-semantic knowledge in the corpora introduced in Section 2, assuming that associations reflect semantic knowledge that can be captured by distributional information. Accordingly, for each of the norms we check the availability of the semantic associates in corpus co-occurrence.

3.1.3 *Study (1):* Association Norms and Corpus Co-Occurrence

Our analyses make use of the following three association norms:

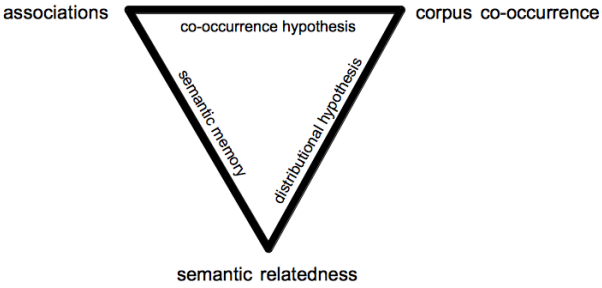

Figure 1: Association–meaning triangle.

1. Associations to German verbs, collected in 2004 (Schulte im Walde et al., 2008):
   - 330 verbs including 36 particle verbs
   - 44–54 participants per stimulus
   - 38,769/79,480 stimulus–association types/tokens

2. Associations to German nouns, collected in 2003/4 (Melinger and Weber, 2006):
   - 409 nouns referring to picturable objects
   - 100 participants per stimulus
   - 30,845/116,714 stimulus–association types/tokens

3. Associations to German noun compounds, collected in 2010–2012:

   a) based on a web experiment with German noun compounds and constituents (Schulte im Walde et al., 2012):
      - 996 compounds+constituents for 442 concrete, depictable compounds
      - 10–36 participants per stimulus
      - 28,238/47,249 stimulus–association types/tokens

   b) based on an Amazon Mechanical Turk (AMT) experiment with a subset of the above compounds+constituents (Borgwaldt and Schulte im Walde, 2013):
      - 571 compounds+constituents for 246 noun-noun compounds
      - 2–120 (mostly: 30) participants per stimulus
      - 26,415/59,444 stimulus–association types/tokens

Relying on these three norms enables us to check on the availability of semantic associate knowledge, with regard to two major word classes of the stimuli (verbs and nouns), and a morphologically restricted subset of one of these classes, compound nouns, where we only take the noun compound stimuli from the norms 3a) and 3b) into account.

Table 3 presents the proportions of the stimulus–associate types that were found in a 5-word window from each other in the various corpora, i.e., the association appeared at least once in a maximum of five words to the left or to the right of the respective stimulus. Since the SdeWaC corpus is sorted alphabetically and thus going beyond the sentence border was likely to entering a sentence that did not originally precede or follow the sentence of interest (cf. Section 2), we distinguish two modes: *sentence-internal (int)*, where window co-occurrence refers to five words to the left and right BUT within the same sentence, and *sentence-external (ext)*, the standard window co-occurrence that goes beyond sentence borders. We applied the *int* mode also to the other corpora, to make the co-occurrence numbers more comparable.

Table 3 shows that the association type coverage is monotonous with regard to the size of the corpora and the co-occurrence mode, with one exception: the larger the corpus, the stronger the coverage of the association data; but for the verb stimuli, the 430-million-word corpus Wiki-de has a lower coverage than the 200-million-word newspaper corpus HGC. Unsurprisingly, including sentence-external co-occurrence material increases the coverage, in comparison to only taking sentence-internal co-occurrence into account.

Norms	Size	Corpora						
		HGC		Wiki-de		WebKo		SdeWaC
		200		*430*		*1,500*		*880*
		ext	int	ext	int	ext	int	int
verbs	38,769	54	51	50	47	73	70	68
nouns	30,845	53	51	56	54	77	76	72
compound nouns	14,326	23	22	25	23	49	47	42

Table 3: Coverage of association types across corpora in a 5-word window.

Table 4 compares the same corpora on association coverage as Table 3, but with regard to sub-corpora of similar sizes, thus looking at the effects of the corpus domains and corpus cleaning. Since the HGC as the smallest corpus contains approx. 200 million words, and accidentally the sizes of the other corpora are rough multiples of 200 million words, we created sub-corpora of approx. 200 million words for each corpus. Table 4 compares the corpus coverage of the verb–association and the noun-association types with regard to the whole HGC, both parts of Wiki-de, two parts (out of eight) of WebKo and all four parts of SdeWaC.

Table 4 shows that the exact coverage of the association types varies from (sub-)corpus to (sub-)corpus but is within the range $[49\%, 57\%]$ for HGC, WebKo and SdeWaC. For Wiki-de, however, the coverage is clearly lower, within the range $[41\%, 49\%]$. Two facts are surprising: (i) We would have expected the association coverage of HGC to be

below that of the web corpora, because the domain is more restricted. The coverage is however compatible with the WebKo and SdeWaC coverage scores. (ii) We would have expected the association coverage of Wiki-de to be compatible with that of the other two web corpora, because the encyclopaedic knowledge was expected to provide relevant semantic knowledge. The coverage is however below that of the other corpora.

Norms	Size	Corpora								
		HGC	Wiki-de		WebKo		SdeWaC			
		ext					int			
verbs	38,769	54	44	41	53	55	56	55	53	49
nouns	30,845	53	49	44	54	56	57	57	55	52

Table 4: Coverage of association types across 200-million-word corpora in a 5-word window.

Altogether, the above two tables demonstrate that the corpora are indeed capturing lexical-semantic knowledge with regard to the various stimulus–association pairs, finding up to 73/77/49% of the types in a small 5-word window of the stimuli in the largest corpus (WebKo). Table 5 (focusing on the SdeWaC) shows that the coverage is even stronger when looking at the stimulus–association pair *tokens* (in comparison to the *types*), reaching 78/87/58% coverage (in comparison to 68/72/42%). Looking at a larger 20-word window, these numbers go up to 84/91/67%. Thus, we induce that we can indeed find lexical-semantic knowledge in our corpora, and stronger related pairs are even more likely to be covered than weaker pairs (which can be induced from the fact that the token coverage is larger than the type coverage).

Norms	Size	SdeWaC			
		window: 5		window: 20	
		types	tokens	types	tokens
verbs	38,769/79,480	68	78	76	84
nouns	30,845/116,714	72	87	80	91
compound nouns	14,326/33,065	42	58	51	67

Table 5: Coverage of association tokens vs. types in *SdeWaC*.

In addition, the larger the corpora are, the stronger is the availability of the semantic knowledge. Since web corpora are the largest corpora available, they therefore represent the best choice for lexical-semantic knowledge, at least with regard to the specific instance of semantic association data. Corpus size outperforms extensive corpus cleaning (when comparing WebKo with SdeWaC in Table 3), but the difference is small (70/76/47% vs. 68/72/42%), taking into consideration that the size difference is large (1,500 vs. 880 million words), so the cleaning obviously had a positive effect on the corpus quality. Finally, general web corpora such as the deWaC (of which WebKo and SdeWaC are

both subsets) seem overall more suitable for general lexical-semantic knowledge than the more structured and entry-based Wikipedia corpus.

Focusing on stimulus–association pairs with compound stimuli in Tables 3 and 5 shows that their coverage is significantly below the coverage of non-compound stimuli (note that most of the stimuli in the verb and noun association norms are not morphologically complex). We assume that this is due to two facts: (i) German noun compounds are notoriously productive, so their corpus coverage is expected to be lower than for non-complex nouns. (ii) Noun compounds are more difficult to process than non-complex nouns because of their morphological complexity, even for standard tools. In sum, we expected a lower co-occurrence corpus coverage for compound–associate pairs in comparison to pairs with morphologically non-complex stimuli, because the stimuli are more sparse. Considering this restriction, the coverage rates are actually quite impressive, reaching 58/67% token coverage in a 5/20-word window of the SdeWaC. Comparing across corpora, the two deWaC corpora capture clearly larger proportions of compound–associate pairs than the HGC and Wiki-de, confirming the usefulness of the two web corpora for the availability of the semantic associate knowledge.

Concerning the stimulus–association pairs that are *not* found in corpus co-occurrence, the reader is referred to Schulte im Walde et al. (2008) and Roth and Schulte im Walde (2008) for detailed explorations. As expected, a serious proportion of these associations reflects world knowledge and is therefore not expected to be found in the immediate context of the stimuli at all, for example *mampfen–lecker* 'munch–yummy', *auftauen–Wasser* 'defrost–water', *Ananas–gelb* 'pineapple–yellow' and *Geschenk–Überraschung* 'present–surprise'. These cases pose a challenge to empirical models of word meaning.

To complete our first analysis on the availability of lexical-semantic knowledge in (web) corpora, Table 6 presents a selection of stimulus–association pairs from different domains, with different morphological complexity of the stimuli, and with difference strengths, accompanied by their 20-word window co-occurrence coverage in our corpora. Comparing WebKo *ext* with WebKo *int* demonstrates that reducing the 20-word windows to sentence-internal windows has a severe effect on the association coverage: all co-occurrence counts for the former condition are larger than for the latter condition, in some cases even twice as much. Furthermore, the table illustrates that typically the larger the corpora the more instances of the stimulus–associate pairs were found. However, the domains of the corpora also play a role: The HGC outperforms Wiki-de in more cases than vice versa, and in four cases (*Affe–Urwald*, *Blockflöte–Musik*, *Polizist–grün*, *Telefonzelle–gelb*), the HGC even outperforms SdeWaC, which is approximately four times as large (however restricted to sentence-internal co-occurrence). Regarding our examples, it is also obvious that the total co-occurrence counts involving compounds are lower than those involving simplex stimuli. Even though the sample is too small to generalise this intuition, it confirms our insight from Table 5 that associations to compounds are covered less than associations to simplex words.

Stimulus–Associate Pairs			Corpus				
			HGC	Wiki-de	WebKo		SdeWaC
				ext	ext	int	int
			200	*430*	*1,500*		*880*
Affe 'monkey'	*Urwald* 'jungle'	15	10	1	89	48	3
analysieren 'analyse'	*untersuchen* 'analyse'	8	56	172	1,613	685	451
bedauern 'regret'	*Mitleid* 'pity'	11	3	0	29	12	10
Blockflöte 'flute'	*Musik* 'music'	23	26	73	90	43	22
Fliegenpilz 'toadstool'	*giftig* 'poisonous'	34	0	9	40	14	9
Kuh 'cow'	*melken* 'milk'	12	71	28	880	683	200
Obstkuchen 'fruit cake'	*backen* 'bake'	7	1	1	17	12	5
Polizist 'police-man'	*grün* 'green'	45	65	9	120	61	52
rollen 'roll'	*Kugel* 'bowl'	15	96	10	654	483	277
schleichen 'crawl'	*leise* 'quiet'	36	10	4	569	428	88
Schlittenhund 'sledge dog'	*Winter* 'winter'	10	1	5	6	3	3
Telefonzelle 'phone box'	*gelb* 'yellow'	25	16	5	17	14	6
verbrennen 'burn'	*heiß* 'hot'	15	42	55	534	348	194

Table 6: Examples of co-occurring stimulus–association pairs across corpora.

3.2 Predicting the Degree of Compositionality of German Noun-Noun Compounds

Our second semantic relatedness task will predict the compositionality of German noun-noun compounds, i.e., the semantic relatedness between a compound and its constituents. The distributional model for this task describes the compound nouns and their nominal constituents by simple corpus co-occurrence features, and relies on the distributional similarity between the compounds and the constituents to rate their semantic relatedness. Section 3.2.1 first introduces the relevant background on German noun compounds and describes our compound data, before Section 3.2.2 presents human ratings on the compositionality of the compounds. Section 3.2.3 then describes the actual distributional explorations.

3.2.1 German Noun-Noun Compounds

Compounds are combinations of two or more simplex words. Traditionally, a number of criteria (such as compounds being syntactically inseparable, and that compounds have a specific stress pattern) have been proposed, in order to establish a border between compounds and non-compounds. However, Lieber and Stekauer (2009a) demonstrated

that none of these tests are universally reliable to distinguish compounds from other types of derived words. Compounds have thus been a recurrent focus of attention within theoretical, cognitive, and in the last decade also within computational linguistics. Recent evidence of this strong interest are the *Handbook of Compounding* on theoretical perspectives (Lieber and Stekauer, 2009b), and a series of workshops and special journal issues on computational perspectives (Journal of Computer Speech and Language, 2005; Language Resources and Evaluation, 2010; ACM Transactions on Speech and Language Processing, 2013).[4]

Our focus of interest is on German noun-noun compounds (see Fleischer and Barz (2012) for a detailed overview and Klos (2011) for a recent detailed exploration), such as *Ahornblatt* 'maple leaf' and *Feuerwerk* 'fireworks', where both the grammatical head (in German, this is the rightmost constituent) and the modifier are nouns. More specifically, we are interested in the degrees of compositionality of German noun-noun compounds, i.e., the semantic relatedness between the meaning of a compound (e.g., *Feuerwerk*) and the meanings of its constituents (e.g., *Feuer* 'fire' and *Werk* 'opus').

Our work is based on a selection of noun compounds by von der Heide and Borgwaldt (2009), who created a set of 450 concrete, depictable German noun compounds according to four compositionality classes: compounds that are transparent with regard to both constituents (e.g., *Ahornblatt* 'maple leaf'); compounds that are opaque with regard to both constituents (e.g., *Löwenzahn* 'lion+tooth → dandelion'); compounds that are transparent with regard to the modifier but opaque with regard to the head (e.g., *Feuerzeug* 'fire+stuff → lighter'); and compounds that are opaque with regard to the modifier but transparent with regard to the head (e.g., *Fliegenpilz* 'fly+mushroom → toadstool'). From the compound set by von der Heide and Borgwaldt, we disregarded noun compounds with more than two constituents (in some cases, the modifier or the head was complex itself) as well as compounds where the modifiers were not nouns. Our final set comprises a subset of their compounds: 244 two-part noun-noun compounds.

3.2.2 Compositionality Ratings

von der Heide and Borgwaldt (2009) collected human ratings on compositionality for all their 450 compounds. The compounds were distributed over 5 lists, and 270 participants judged the degree of compositionality of the compounds with respect to their first as well as their second constituent, on a scale between 1 (definitely opaque) and 7 (definitely transparent). For each compound–constituent pair, they collected judgements from 30 participants, and calculated the rating mean and the standard deviation.

Table 7 presents example mean ratings for the compound–constituent ratings, accompanied by the standard deviations. We selected two examples each from our set of 244 noun-noun compounds, according to five categories of mean ratings: the compound–constituent ratings were (1) high or (2) mid or (3) low with regard to both constituents; the compound–constituent ratings were (4) low with regard to the modifier but high with regard to the head; (5) vice versa.

[4] www.multiword.sourceforge.net

Compounds			Mean Ratings	
whole	literal meanings of constituents		modifier	head
Ahornblatt 'maple leaf'	maple	leaf	**5.64** ± 1.63	**5.71** ± 1.70
Postbote 'post man'	mail	messenger	**5.87** ± 1.55	**5.10** ± 1.99
Seezunge 'sole'	sea	tongue	**3.57** ± 2.42	**3.27** ± 2.32
Windlicht 'storm lamp'	wind	light	**3.07** ± 2.12	**4.27** ± 2.36
Löwenzahn 'dandelion'	lion	tooth	**2.10** ± 1.84	**2.23** ± 1.92
Maulwurf 'mole'	mouth	throw	**2.21** ± 1.68	**2.76** ± 2.10
Fliegenpilz 'toadstool'	fly/bow tie	mushroom	**1.93** ± 1.28	**6.55** ± 0.63
Flohmarkt 'flea market'	flea	market	**1.50** ± 1.22	**6.03** ± 1.50
Feuerzeug 'lighter'	fire	stuff	**5.87** ± 1.01	**1.90** ± 1.03
Fleischwolf 'meat chopper'	meat	wolf	**6.00** ± 1.44	**1.90** ± 1.42

Table 7: Examples of compound ratings.

3.2.3 *Study (2):* Predicting Compositionality by Similarity of Corpus Co-Occurrence

In this study, the goal of our experiments is to predict the degree of compositionality of our set of noun-noun compounds as presented in the previous section, by relying on the similarities between the compound and constituent distributional properties. The distributional properties of our noun targets (i.e., the compounds as well as the nominal constituents) are instantiated by a standard vector space model (Turney and Pantel, 2010; Erk, 2012) using window co-occurrence with varying window sizes. We restrict window co-occurrence to nouns, i.e., the dimensions in the vector spaces are all nouns co-occurring with our target nouns within the specified window sizes.[5] For example, for a window size of 5, we count how often our target nouns appeared with any nouns in a window of five words to the left and to the right. As in our first case study, we distinguish the two modes *sentence-internal (int)* and *sentence-external (ext)*.

In all our vector space experiments, we first induce co-occurrence frequency counts from our corpora, and then calculate *local mutual information (LMI)* values (Evert, 2005), to instantiate the empirical properties of our target nouns. LMI is a measure from information theory that compares the observed frequencies O with expected frequencies E, taking marginal frequencies into account: $LMI = O \times log \frac{O}{E}$, with E representing the product of the marginal frequencies over the sample size.[6] In comparison to (pointwise) mutual information (Church and Hanks, 1990), LMI improves the problem of propagating low-frequent events.

Relying on the LMI vector space models, the *cosine* determines the distributional similarity between the compounds and their constituents, which is in turn used to predict the compositionality between the compound and the constituents, assuming that the stronger the distributional similarity (i.e., the *cosine* values), the larger the degree of compositionality. The vector space predictions are evaluated against the human ratings on the degree of compositionality (cf. Section 3.2.2), using the Spearman

[5] See Schulte im Walde et al. (2013) for variations of this noun vector space.

[6] See `http://www.collocations.de/AM/` for a detailed illustration of association measures.

Rank-Order Correlation Coefficient ρ (Siegel and Castellan, 1988). The ρ correlation is a non-parametric statistical test that measures the association between two variables that are ranked in two ordered series.

Figure 2 presents the correlation coefficient ρ (i.e., the quality of the predictions) for our three web corpora WebKo, SdeWaC and Wiki-de across the window sizes $1, 2, 5, 10, 20$. As in our first study, WebKo once more outperforms the SdeWaC corpus, for both modes *ext* and *int*, reaching an optimal prediction of $\rho = 0.6497$ (WebKo (ext)) when relying on a 20-word noun window. For larger windows, the difference between WebKo (ext) and SdeWaC is even significant, according to the Fisher r-to-z transformation. The difference between WebKo (int) and SdeWaC is marginal, however, especially taking into account that WebKo is twice as big as SdeWaC. As in our previous study, the task performance relying on Wiki-de is significantly worse than when relying on WebKo or SdeWaC.

As none of the window lines in Figure 2 has reached an optimal correlation with a window size of 20 yet (i.e., the correlation values are still increasing), we enlarged the window size up to 100 words, in order to check on the most successful window size. For SdeWaC and Wiki-de, the correlations slighly increase, but for WebKo they do not increase. The optimal prediction is still performed using WebKo (ext) and a window size of 20 (see above).

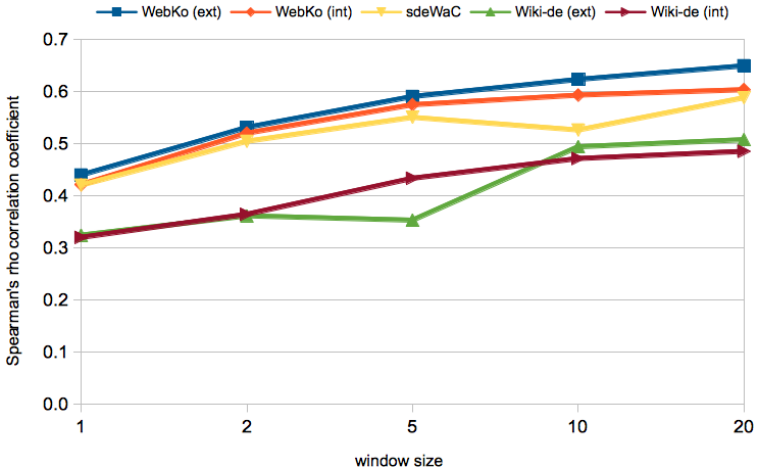

Figure 2: Window-based ρ correlations across corpora.

Figure 3 compares the three corpora on the same task, prediction of compositionality, but with regard to sub-corpora of similar sizes, thus looking at the effects of the corpus domains and corpus cleaning. The left-hand part of the plot breakes WebKo down into two sub-corpora of approx. 800 million words, and compares the prediction quality with

SdeWaC. This part of the plot once more confirms that "real" windows going beyond the sentence border clearly outperform window information restricted to the sentence level, reaching ρ correlations of 0.6494/0.6265 (WebKo (ext)) in comparison to 0.6048/0.5820 (WebKo (int)). The difference between the directly comparable WebKo (int) and SdeWaC vanishes, in contrast, as the WebKo (int) sub-corpora do not consistently outperform the SdeWaC ρ correlation of 0.5883. So in this study the corpus cleaning does not have an obvious effect on the semantic task.

The right-hand part of the plot breakes down SdeWaC into two sub-corpora of approx. 440 million words, and compares the prediction quality with Wiki-de (in both window modes). In this case, the difference in prediction quality persists: SdeWaC does not only outperform the directly comparable Wiki-de (int), $\rho = 0.4857$, but also Wiki-de (ext), $\rho = 0.5080$, which is in an advantegeous position, as Wiki-de (ext) is roughly of the same size as SdeWaC but window-based beyond the sentence border. So again, the web corpus Wiki-de performs worse than both other web corpora.

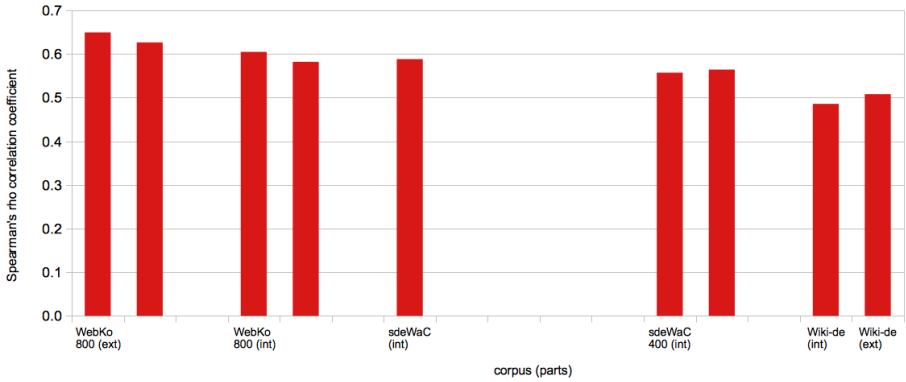

Figure 3: Window-based ρ correlations across corpus parts.

4 Summary and Discussion

The previous section presented two case studies to explore whether and how web corpora can be used to automatically acquire lexical-semantic knowledge from distributional information. For this purpose, we compared three German web corpora and one newspaper corpus, differing in domain, size and cleanness. Both case studies demonstrated that the corpora can indeed be used to model semantic relatedness: In the first study, we found up to 73/77% of verb/noun–association types within a 5-word window.[7]

[7] Schulte im Walde et al. (2008), Schulte im Walde and Melinger (2008) and Roth and Schulte im Walde (2008) present related work on co-occurrence analyses of the association norms.

Adhering to the standard assumption that *associates reflect meaning components of words*, we thus presented strong evidence for the *co-occurrence hypothesis* as well as for the *distributional hypothesis*, that semantic associations can be captured by corpus co-occurrence. In the second study, we could predict the compositionality of German noun-noun compounds with regard to their constituents with a ρ correlation of up to 0.6497.[8] We were therefore successful in inducing semantic compositionality, based on the distributional information in the web corpora.

Comparing the web corpora with a standard newspaper corpus (only done in the first study), we found that the availability of association knowledge in the newspaper corpus was actually compatible with the availability in the web corpora. So there was no effect concerning the more restricted domain, even though the association knowledge is completely open-domain.

In contrast, we found an effect of web corpus size in both studies: the larger the corpus, (1) the stronger the coverage of the association data, and (2) the better the prediction of compositionality. The differences between the larger (and noisier) WebKo data and the smaller (and cleaner) SdeWaC data were small (in study (1)) and negligible (in study (2)), indicating that the corpus cleaning had a positive effect on the corpus quality. Comparing the general web corpora WebKo and SdeWaC with the more structured and entry-based Wikipedia corpus, we demonstrated that the former are more suitable both (1) for general semantic knowledge and also (2) for predicting noun compound compositionality. We would have expected the encyclopeadic knowledge in Wiki-de to be compatible with the association knowledge in task (1) but the coverage is below that of the other two web corpora, even if we take the size into account.

Summarising, this article confirmed the suitability of web corpora for the automatic acquisition of lexical-semantic knowledge with regard to two very diverse semantic case studies. At the same time, we showed that newspaper corpora – if equal in size – provide suffient semantic relatedness knowledge, too, and are thus less restricted in their domains than commonly assumed.

Acknowledgements

The research presented in this article was funded by the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde) and the DFG Sachbeihilfe SCHU-2580/2-1 (Stefan Müller). In addition, we thank the anonymous reviewer for valuable comments.

[8] Schulte im Walde et al. (2013) present related work on predicting the degree of compositionality of the noun-noun compounds.

References

Banko, M. and Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Toulouse, France.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M. and Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4):673–721.

Boleda, G., Baroni, M., The Pham, N., and McNally, L. (2013). Intensionality was only alleged: On Adjective-Noun Composition in Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 35–46, Potsdam, Germany.

Borgwaldt, S. and Schulte im Walde, S. (2013). A Collection of Compound–Constituent and Compound Whole Ratings for German Noun Compounds. Manuscript.

Church, K. W. and Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, Canada.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Clark, H. H. (1971). Word Associations and Linguistic Theory. In Lyons, J., editor, *New Horizon in Linguistics*, chapter 15, pages 271–286. Penguin.

Curran, J. and Moens, M. (2002). Improvements in Automatic Thesaurus Extraction. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, pages 59–66, Philadelphia, PA.

de Deyne, S. and Storms, G. (2008a). Word Associations: Network and Semantic Properties. *Behavior Research Methods*, 40(1):213–231.

de Deyne, S. and Storms, G. (2008b). Word Associations: Norms for 1,424 Dutch Words in a Continuous Task. *Behavior Research Methods*, 40(1):198–205.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6(10):635–653.

Erk, K., Padó, S., and Padó, U. (2010). A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–763.

Evert, S. (2005). *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Faaß, G. and Eckart, K. (2013). SdeWaC – a Corpus of Parsable Sentences from the Web. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pages 61–68, Darmstadt, Germany.

Faaß, G., Heid, U., and Schmid, H. (2010). Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 803–810, Valletta, Malta.

Fernández, A., Diez, E., Alonso, M. A., and Beato, M. S. (2004). Free-Association Norms for the Spanish Names of the Snodgrass and Vanderwart Pictures. *Behavior Research Methods, Instruments and Computers*, 36(3):577–583.

Ferrand, L. and Alario, F.-X. (1998). French Word Association Norms for 366 Names of Objects. *L'Ann'ee Psychologique*, 98(4):659–709.

Firth, J. R. (1957). *Papers in Linguistics 1934-51*. Longmans, London, UK.

Fleischer, W. and Barz, I. (2012). *Wortbildung der deutschen Gegenwartssprache*. de Gruyter.

Glenberg, A. M. and Mehta, S. (2008). Constraint on Covariation: It's not Meaning. *Italian Journal of Linguistics. Alessandro Lenci (guest editor): From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science*, 20(1):241–264.

Harris, Z. (1968). Distributional Structure. In Katz, J. J., editor, *The Philosophy of Linguistics*, Oxford Readings in Philosophy, pages 26–47. Oxford University Press.

Herdagdelen, A. and Baroni, M. (2009). BagPack: A General Framework to Represent Semantic Relations. In *Proceedings of the EACL Workshop on Geometrical Models for Natural Language Semantics*, pages 33–40, Athens, Greece.

Heringer, H. J. (1986). The Verb and its Semantic Power: Association as the Basis for Valence. *Journal of Semantics*, 4:79–99.

Hickl, A., Williams, J., Bensley, J., Roberts, K., Rink, B., and Shi, Y. (2006). Recognizing Textual Entailment with LCC's GROUNDHOG System. In *Proceedings of the Second PASCAL Challanges Workshop on Recognising Textual Entailment*, pages 80–85, Venice, Italy.

Hirsh, K. W. and Tree, J. (2001). Word Association Norms for two Cohorts of British Adults. *Journal of Neurolinguistics*, 14(1):1–44.

Kiss, G. R., Armstrong, C., Milroy, R., and Piper, J. (1973). An Associative Thesaurus of English and its Computer Analysis. In *The Computer and Literary Studies*. Edinburgh University Press.

Klos, V. (2011). *Komposition und Kompositionalität*. Number 292 in Germanistische Linguistik. Walter de Gruyter, Berlin.

Lauteslager, M., Schaap, T., and Schievels, D. (1986). *Schriftelijke Woordassociatienormen voor 549 Nederlandse Zelfstandige Naamworden*. Swets and Zeitlinger.

Lieber, R. and Stekauer, P. (2009a). Introduction: Status and Definition of Compounding. In Lieber and Stekauer (2009b), chapter 1, pages 3–18.

Lieber, R. and Stekauer, P., editors (2009b). *The Oxford Handbook of Compounding*. Oxford University Press.

Marconi, D. (1997). *Lexical Competence*. MIT Press, Cambridge, MA.

McCarthy, D., Keller, B., and Carroll, J. (2003). Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding Predominant Senses in Untagged Text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain.

McKoon, G. and Ratcliff, R. (1992). Spreading Activation versus Compound Cue Accounts of Priming: Mediated Priming Revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18:1155–1172.

McNamara, T. P. (2005). *Semantic Priming: Perspectives from Memory and Word Recognition*. Psychology Press, New York.

Melinger, A. and Weber, A. (2006). Database of Noun Associations for German. URL: www.coli.uni-saarland.de/projects/nag/.

Miller, G. (1969). The Organization of Lexical Memory: Are Word Associations sufficient? In Talland, G. A. and Waugh, N. C., editors, *The Pathology of Memory*, pages 223–237. Academic Press, New York.

Nelson, D. L., Bennett, D., and Leibert, T. (1997). One Step is not Enough: Making Better Use of Association Norms to Predict Cued Recall. *Memory and Cognition*, 25:785–796.

Nelson, D. L., McEvoy, C. L., and Dennis, S. (2000). What is Free Association and What does it Measure? *Memory and Cognition*, 28:887–899.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). The University of South Florida Word Association, Rhyme, and Word Fragment Norms. http://www.usf.edu/FreeAssociation/.

Padó, S. and Utt, J. (2012). A Distributional Memory for German. In *Proceedings of the 11th Conference on Natural Language Processing*, pages 462–470, Vienna, Austria.

Palermo, D. and Jenkins, J. J. (1964). *Word Association Norms: Grade School through College*. University of Minnesota Press, Minneapolis.

Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards Terascale Knowledge Acquisition. In *Proceedings of the 20th International Conference of Computational Linguistics*, pages 771–777, Geneva, Switzerland.

Plaut, D. C. (1995). Semantic and Associative Priming in a Distributed Attractor Network. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, volume 17, pages 37–42.

Quasthoff, U., Richter, M., and Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1799–1802, Genoa, Italy.

Rapp, R. (2002). The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Reddy, S., McCarthy, D., and Manandhar, S. (2011). An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Roth, M. and Schulte im Walde, S. (2008). Corpus Co-Occurrence, Dictionary and Wikipedia Entries as Resources for Semantic Relatedness Information. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1852–1859, Marrakech, Morocco.

Russell, W. A. (1970). The complete German Language Norms for Responses to 100 Words from the Kent-Rosanoff Word Association Test. In Postman, L. and Keppel, G., editors, *Norms of Word Association*, pages 53–94. Academic Press, New York.

Russell, W. A. and Meseck, O. (1959). Der Einfluss der Assoziation auf das Erinnern von Worten in der deutschen, französischen und englischen Sprache. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 6:191–211.

Schiehlen, M. (2003). A Cascaded Finite-State Parser for German. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 163–166, Budapest, Hungary.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging using Decision Trees. In *Proceedings of the 1st International Conference on New Methods in Language Processing*.

Schulte im Walde, S. (2008). Human Associations and the Choice of Features for Semantic Verb Classification. *Research on Language and Computation*, 6(1):79–111.

Schulte im Walde, S. (2010). Comparing Computational Approaches to Selectional Preferences: Second-Order Co-Occurrence vs. Latent Semantic Clusters. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1381–1388, Valletta, Malta.

Schulte im Walde, S., Borgwaldt, S., and Jauch, R. (2012). Association Norms of German Noun Compounds. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 632–639, Istanbul, Turkey.

Schulte im Walde, S. and Melinger, A. (2008). An In-Depth Look into the Co-Occurrence Distribution of Semantic Associates. *Italian Journal of Linguistics. Alessandro Lenci (guest editor): From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science*, 20(1):89–128.

Schulte im Walde, S., Melinger, A., Roth, M., and Weber, A. (2008). An Empirical Characterisation of Response Types in German Association Norms. *Research on Language and Computation*, 6(2):205–238.

Schulte im Walde, S., Müller, S., and Roller, S. (2013). Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, GA.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123. Special Issue on Word Sense Disambiguation.

Siegel, S. and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, MA.

Spence, D. P. and Owens, K. C. (1990). Lexical Co-Occurrence and Association Strength. *Journal of Psycholinguistic Research*, 19:317–330.

Springorum, S., Schulte im Walde, S., and Utt, J. (2013). Detecting Polysemy in Hard and Soft Cluster Analyses of German Preposition Vector Spaces. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 632–640, Nagoya, Japan.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

von der Heide, C. and Borgwaldt, S. (2009). Assoziationen zu Unter-, Basis- und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1646–1652, Marrakech, Morocco.