Corpus Co-Occurrence, Dictionary and Wikipedia Entries as Resources for Semantic Relatedness Information

Michael Roth*, Sabine Schulte im Walde**
*Computational Linguistics, Saarland University, Saarbrücken, Germany
**Institute for Natural Language Processing, University of Stuttgart, Stuttgart, Germany
mroth@coli.uni-sb.de, schulte@ims.uni-stuttgart.de

Abstract

Distributional, corpus-based descriptions have frequently been applied to model aspects of word meaning. However, distributional models that use corpus data as their basis have one well-known disadvantage: Even though the distributional features based on corpus co-occurrence were often successful in capturing meaning aspects of the words to be described, they generally fail to capture those meaning aspects that refer to world knowledge, because coherent texts tend not to provide redundant information that is presumably available knowledge. The question we ask in this paper is whether dictionary and encyclopaedic resources might complement the distributional information in corpus data, and provide world knowledge that is missing in corpora. As test case for meaning aspects, we rely on a collection of semantic associates to German verbs and nouns. Our results indicate that a combination of the knowledge resources should be helpful in work on distributional descriptions.

1. Motivation

Research in data-intensive lexical semantics aims to empirically define and induce features that (a) capture the various meaning aspects of the words to be described, and (b) one can obtain automatically; the goal is to determine the meaning as well as the similarity or dissimilarity of words, sentences, paragraphs, or even documents. Following the distributional hypothesis, namely that 'each language can be described in terms of a distributional structure, i.e., in terms of the occurrence of parts relative to other parts' (Harris, 1968), distributional, corpus-based descriptions have frequently been applied to model aspects of word meaning. For example, variants of the vector space model (Salton et al., 1975) that uses words in documents to describe the contents of the respective documents, have been used in various NLP tasks and applications including word sense discrimination (Schütze, 1998), anaphora resolution (Poesio et al., 2002), general models of semantic similarity (Lin, 1998; Schulte im Walde, 2006; Sahlgren, 2006; Padó and Lapata, 2007), and also in psycholinguistic models of semantic priming (Lund et al., 1995; Lowe and McDonald, 2000; Vigliocco et al., 2004).

Distributional models that use corpus data as their basis have one well-known disadvantage: Even though the distributional features based on corpus co-occurrence were often successful in capturing meaning aspects of the words to be described, they generally fail to capture those meaning aspects that refer to world knowledge, because coherent texts tend not to provide redundant information that is presumably available knowledge. This fact has recently been illustrated by Schulte im Walde et al. (to appear): A collection of semantic associates to German verbs and nouns was used as a test case for covering meaning aspects by corpus co-occurrence (among other analyses).[1] The assumption was that the evoked words, i.e., the associates, reflect highly salient linguistic and conceptual features of the respective stimulus words. Among other analyses, Schulte im Walde et al. demonstrated that corpus co-occurrence did only account for 67% of the associate types to verbs and for 70% of the associate types to nouns. The remaining 30-33% were partly due to lemmatisation mismatches and the domain (newspaper) and size of the corpus. However, further cases of associations that did not appear in co-occurrence with their stimulus words reflected world knowledge, and were therefore unlikely to be found in the immediate context of the stimuli at all (e.g., *Wasser* 'water' for *auftauen* 'defrost', *Überraschung* 'surprise' for *Geschenk* 'present', and *gelb* 'yellow' for *Ananas* 'pineapple').

The question we ask in this paper is whether dictionary and encyclopaedic resources might complement the distributional information in corpus data, and provide world knowledge that is missing in corpora. Dictionary and encyclopaedic knowledge resources have a long history as semantic resources, from the early days in machine translation (Bar-Hillel, 1960), approaches to word sense disambiguation (Lesk, 1986), and more recently –relying on Wikipedia– in NLP tasks such as text categorisation (Gabrilovich and Markovitch, 2006), word sense disambiguation (Mihalcea, 2007), and co-reference resolution (Ponzetto and Strube, 2007). These resources provide detailed information about word senses, and include world knowledge to varying degrees. (Cf., for example, Weber (1996) and Engelberg and Lemnitzer (2004) about dictionary and encyclopaedic information.) We rely on the same association norms that have been used by Schulte im Walde et al., and check on how many and which of the semantic associates are found within the dictionary/encyclopaedic entries of the respective stimulus words. We compare the results against the previous co-occurrence analyses, to decide whether the different types of information complement each other. The positive outcome is interesting for work on distributional descriptions, by indicating resources that potentially add knowledge to word meaning descriptions.

[1] *Semantic associates* are defined here as words that are spontaneously called to mind by a stimulus word.

2. Association norms

This section introduces our methods for collecting human associations to German verbs and nouns,[2] and a distributional representation of the data as stimulus-associate type frequencies. For more details on the data collection, the reader is referred to Schulte im Walde et al. (to appear).

2.1 Associates to verb stimuli

The data collection of associates to verb stimuli was performed as a web experiment, which asked native speakers to provide associations to German verbs. 330 verbs were selected for the experiment, drawn from a variety of semantic classes and from different corpus frequency ranges. Each collection trial consisted of a verb presented in a box at the top of the screen. Below the verb was a series of data input lines where participants could type their associations. Participants had 30 seconds per verb to type as many associations as they could. In total, we collected 79,480 associate responses distributed over 39,254 different response types. Table 1 provides an example of the token-per-type association frequencies, listing the 10 most frequent responses for the polysemous verb *klagen* 'complain, moan, sue'.

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

Table 1: Associate frequencies for stimulus verb.

2.2 Associates to noun stimuli

The data collection of associates to noun stimuli was performed as an offline experiment (Melinger and Weber 2006),[3] which asked native speakers to provide up to three associations to German nouns. 409 German nouns referring to picturable objects were chosen as target stimuli, again drawn from a variety of semantic classes and from different corpus frequency ranges. No time limits were given for responding, though participants were told to work swiftly and without interruption. In total, they collected 116,714 associate responses distributed over 31,035 different response types. Table 2 provides an example of the token-per-type association frequencies, listing the 10 most frequent responses for the polysemous noun *Schloss* 'lock, castle'.

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 frequencies for stimulus noun.

3. Knowledge resources

This section introduces the resources that contributed to the characterisation of the association norms.

3.1 Corpus data

A German newspaper corpus from the 1990s was used for co-occurrence analyses between verb/noun stimuli and associate responses. The corpus contains approximately 200 million words of newspaper text.

3.2 Dictionary: WDG

The dictionary-based analysis relies on the freely available "Wörterbuch der Deutschen Gegenwartssprache" (WDG).[4] The online resource WDG consists of 130,000 entries. In each entry, there is a detailed description of senses, idioms and compounds containing the respective word. For our analyses, we downloaded all entries for our verb and noun stimulus words. This resulted in 706 entries (for 33 stimulus words, the dictionary did not provide any entry), containing a total of 51,561/593,318 words (types/token), with an average of 330/840 words (types/token) per entry. Table 3 provides as example a part of the WDG entry for the ambiguous noun *Apfel* 'apple', referring to two senses of *Apfel,* 'fruit of apple tree' and 'apple tree (coll.)', each followed by example contexts (in italics). Morphological information is omitted in the example.

Apfel 'apple'
1. Frucht des Apfelbaums: *ein reifer rotbackiger, grüner, runder, kandierter, gebratener, geriebener, fauler, schrumpeliger, wurmstichiger Apfel* (...)
2. (umg.) Apfelbaum: *dieser Apfel blüht früh, trägt besonders gut*
...

Table 3: Example dictionary entry.

3.2 Encyclopaedia: Wikipedia

Wikipedia is a free multi-lingual online encyclopaedia.[5]

[2] The association norms for verbs and nouns were collected in independent studies with different goals; as a consequence they differ somewhat in the methods used for data collection.

[3] http://www.coli.uni-saarland.de/projects/nag/

[4] http://www.dwds.de

[5] http://www.wikipedia.org

The resource represents the result of a collaborative effort, where anybody can create and edit entries. This paper uses the German version of Wikipedia, which contained over 650,000 entries as of October 2006. For our analyses, we downloaded all Wikipedia articles which were labelled by any of the verb or noun stimulus words. If the downloaded article indicated an ambiguity, we also downloaded the linked pages, thus using all available articles for the various senses of the respective word. This resulted in 2,447 articles for 542 stimulus words (for 197 stimulus words, Wikipedia did not provide any entry), containing a total of 270,014/2,848,150 words (types/token), with an average of 492/1,164 words (types/token) per article. Table 4 provides as example a part of the Wikipedia linked page for the ambiguous noun *Apfel* 'apple', and the beginnings of the respective Wikipedia articles for the two linked senses *Äpfel* and *Kulturapfel*.

Apfel 'apple'
Apfel steht für 'apple stands for': 1. Äpfel, allgemein einen Baum der Gattung Malus 2. im engeren Sinn einen Baum der Art Malus domestica oder dessen Frucht, siehe Kulturapfel und Liste der Apfelsorten ...
1. *Äpfel* 'malus': Die Äpfel (Malus) bilden eine Gattung in der Unterfamilie der Kernobstgewächse (Maloideae) aus der Familie der Rosengewächse (Rosaceae). (...)
2. *Kulturapfel* 'apple': Der Kulturapfel oder auch kurz Apfel (Malus domestica) ist eine weithin bekannte Art aus der Gattung der Äpfel in der Familie der Rosengewächse (Rosaceae). (...)

Table 4: Example Wikipedia linked page and articles.

4. Analyses

The analyses check for each stimulus-associate pair (token and type), whether it is covered by the knowledge sources introduced in the previous Section 3.

4.1 Corpus-based analysis

This analysis is taken from Schulte im Walde et al. (to appear). We checked whether the stimulus-associate pairs co-occurred in our corpus within a window of 20 words to the left and to the right of each other. [6] Table 5 and Table 6 present the results. The *'all'* row shows the percentage of associate responses that were found in co-occurrence with their stimulus words across all parts-of-speech (POS) of the responses; the following rows are broken down for the parts-of-speech of the responses. Both for the verb and noun stimuli, 1% of the words were missing in the corpus and therefore a priori missing in the analyses.

POS	Types	Tokens
all	***70%***	***84%***
N	69%	84%
V	76%	88%
ADJ	72%	83%

Table 5: Noun-association co-occurrences.

POS	Types	Tokens
all	***67%***	***77%***
N	66%	76%
V	67%	79%
ADJ	70%	77%
ADV	89%	91%

Table 6: Verb-association co-occurrences.

4.2 Dictionary-based analysis

Table 7 and Table 8 present the token and type coverage of the stimulus-associate pairs in the dictionary. The *'all'* row shows the percentage of associate responses that were found in the dictionary entries of the respective stimulus words across all POS of the responses (at least once); the following rows are broken down for the parts-of-speech of the responses. The column 'missing' shows how many pairs (types/tokens) could not be processed because there was no entry for the stimulus word.

POS	Types	Tokens	missing
all	***12%***	***28%***	***7% / 7%***
N	11%	26%	7% / 7%
V	21%	43%	7% / 5%
ADJ	16%	31%	8% / 8%

Table 7: Noun-association coverage by dictionary.

POS	Types	Tokens	missing
all	***13%***	***26%***	***0% / 0%***
N	12%	25%	0% / 0%
V	16%	29%	0% / 0%
ADJ	13%	24%	0% / 0%
ADV	23%	28%	1% / 1%

Table 8: Verb-association coverage by dictionary.

4.3 Encyclopaedia-based analysis

Table 9 and Table 10 present the respective results of the stimulus-associate pairs in Wikipedia.

POS	Types	Tokens	missing
all	***26%***	***46%***	***2% / 2%***
N	28%	49%	2% / 2%
V	22%	37%	2% / 2%
ADJ	24%	39%	2% / 3%

Table 9: Noun-association coverage by Wikipedia.

[6] We are aware of the fact that a co-occurrence strength of 1 does not provide any strong evidence for the relatedness of the stimuli and the responses. The original analyses in Schulte im Walde et al. therefore broke the results down to several co-occurrence strengths. In this article, however, we rely on a strength of 1 in accordance with the other analyses.

POS	Types	Tokens	missing
all	***6%***	***10%***	***56% / 54%***
N	7%	12%	54% / 52%
V	4%	6%	58% / 58%
ADJ	6%	9%	58% / 54%
ADV	10%	9%	55% / 60%

Table 10: Verb-association coverage by Wikipedia.

4.4 Comparison and interpretation

This section contains the main body of our work. We compare the coverage of the three resources (Section 4.4.1), the type of knowledge that is covered (Section 4.4.2), and discuss the role of frequency for the coverage (Section 4.4.3).

4.4.1 Comparison of resource coverage

Comparing the overall *'all'* proportions of stimulus-association pairs that are covered by the three knowledge resources shows that the corpus data covers more pairs than Wikipedia, which in turn covers more than the dictionary. An important factor regarding the coverage is, of course, the size of the various resources, which we will address in Section 4.4.3. For the moment, we concentrate on comparisons that are within-resource: The resources differ in

a. the proportions of coverage and missing entries when comparing associates to nouns vs. verbs,
b. the coverage with respect to the parts-of-speech of the responses,
c. the token-type ratio of the covered stimulus-association pairs.

Concerning a. – ***The resources differ in the parts-of-speech of the stimuli about which they provide knowledge.*** While the corpus is missing 1% of both verb and noun stimuli (about which it consequently cannot provide knowledge), the dictionary covers all verb-association pairs but fails to cover 7% of the noun-association pairs because the respective noun stimulus entries were missing; Wikipedia is missing only 2% of the noun-association pairs but 56% of the verb-association pairs, because the respective stimulus entries were missing. Accordingly, the overall coverage of the resources with respect to noun vs. verb stimuli is similar for the corpus co-occurrence (70/67%) and the dictionary (12/13%), but differs strongly for Wikipedia (26/6%). Whether these differences represent a general tendency of the resource types, and whether they apply across to other, similar resources (e.g., to other dictionaries, and other encyclopaedias) remains an open question within this paper but is certainly an interesting issue to address, as it might provide insight into preferences for resource types according to the parts-of-speech of words that are described by the respective resource knowledge.

Concerning b. – ***The resources differ in the proportions of the various parts-of-speech of the associations that they cover.*** With respect to *association types to noun stimuli,* Wikipedia covers more noun responses than adjective responses, and more adjective responses than verb responses (N > ADJ > V); for corpus co-occurrence and the dictionary, in contrast, the proportions are related as V > ADJ > N. With respect to *association types to verb stimuli,* all resources provide a strong coverage of adverb responses; the other parts-of-speech of the responses are covered similarly strong within each resource, with small differences in the proportions. Comparing the coverage of the various parts-of-speech, of course, needs to be related to the overall numbers of the parts-of-speech within the resources. For example, Wikipedia contains a total of 192,655/883,059 nouns (types/tokens), 19,527/260,156 verbs, and 41,164/238,475 adjectives; WDG contains a total of 22,979/135,113 nouns (types/tokens), 10,375/69,269 verbs, and 11,602/48,471 adjectives. This comparison shows, however, that the proportions of the associate parts-of-speech that are found within the resources are not correlated with the overall part-of-speech proportions in the resources; thus, we can infer that there must be other, resource-based reasons to the different patterns of part-of-speech coverage. The high coverage of adverbs in all three resources can be explained by the token-type ratio of adverbs: even though adverbs are an open class in German, the absolute number of adverb types in natural language text and speech is much lower than those of verbs, nouns, and adjectives, but at the same time the token-per-type ratio is higher (for example, Wikipedia contains 861/92,321 adverbs (types/tokens), and the dictionary contains 452/13,185 adverbs); furthermore, the grammatical restrictions within German clause structure are lower. So there is a high prior probability to find adverbs in the vicinity of a verb.

Concerning c. – ***The resources differ in the token-type ratio of their overall coverage.*** Comparing the *'all'* proportions for token vs. type coverage gives a factor of 1.2/1.2 for the noun-association and the verb-association pairs covered by corpus co-occurrence; 1.8/1.7 in Wikipedia, and 2.3/2.0 in the dictionary. The higher the ratio is, the more of the stronger associate responses were among the ones that were found in the respective resources. We conclude, that –even though the dictionary has a lower overall coverage of stimulus-associate pairs than the other two resources– it covers rather strongly associated responses, and its knowledge therefore might be considered as a better fit to association data than the other resources. Wikipedia, then, has in turn a better fit than corpus co-occurrence.

4.4.2 Comparison of resource knowledge

Table 11 and Table 12 address our key question, whether dictionary and encyclopaedic knowledge sources complement corpus co-occurrence and thus standard distributional analyses. As a first step towards this question, the tables provide a cross-comparison of the resource coverage with respect to the three resources. The tables show how many of the stimulus-associate pairs were covered by the resource in the first column vs. the other resources listed in columns 2-4. For example, concerning

the noun-association types, corpus co-occurrence covered 55% more of the stimulus-associate pairs than the dictionary, and 46% more than Wikipedia. The tables complement the tables in Section 4.1-4.3, also showing that the corpus coverage is larger than the dictionary and Wikipedia coverage; in addition, the tables show that each resource covers some stimulus-associate pairs that the other resources do not cover, to varying degrees. The proportions in the tables refer to the stimulus-associate *types*, as we are interested in the kind of information that is provided by the knowledge resources.

	Corpus	Dic	Wiki
Corpus	-	55.0%	46.0%
Dic	0.8%	-	5.7%
Wiki	3.2%	18.1%	-

Table 11: Cross-comparison of resource coverage of noun-association types.

	Corpus	Dic	Wiki
Corpus	-	45.8%	22.1%
Dic	0.7%	-	3.9%
Wiki	0.5%	3.6%	-

Table 12: Cross-comparison of resource coverage of verb-association types.

To illustrate the differences in knowledge that is covered by the three resources, we rely on examples, in relation to the cross-comparison tables. The appendix in Section 7 lists the 10 strongest stimulus-associate pairs for each cross-comparison. Following a descriptive approach, we find stimulus-associate pairs that might indeed be considered as world knowledge and are covered by Wikipedia or the dictionary (but not by corpus co-occurrence), such as *Karotte* 'carrot' – *orange* 'orange', *Zebra* 'zebra' – *Streifen* 'stripes', *Reibe* 'grater' – *Käse* 'cheese', *auftauen* 'defrost' – *Wärme* 'heat', *einfrieren* 'freeze' – *Kühlschrank* 'fridge'. Similarly, however, we find pairs that might be considered as world knowledge and are covered by corpus co-occurrence but not by the dictionary or Wikipedia, such as *Iglu* 'igloo' – *Eskimo* 'Eskimo', *Stecker* 'connector' – *Strom* 'electricity', *Zitrone* 'lemon' – *sauer* 'sour', *donnern* 'thunder' – *Gewitter* 'thunderstorm', *lehren* 'teach' – *Schule* 'school'. These examples are based on intuitions only but nevertheless illustrate that in sum, we cannot conclude from the examples that dictionary and encyclopaedic resources provide more world knowledge than corpus co-occurrence. In order to distinguish the various knowledge types on a more general basis, one would have to work with word pairs (or other data) that are annotated for their semantic relation. Nevertheless, Tables 11-12 and the examples allow us to conclude that the knowledge in the various resources complements each other, and that combining the knowledge that is provided by corpus co-occurrence, dictionaries and encyclopaedias might therefore enhance distributional descriptions of words.

4.4.3 Taking size and frequencies into account

A central concern in the interpretation of the results, which we have not addressed so far, is of course the various sizes of the resources and the individual entries. In this section, we discuss the potential objections the reader might have concerning the resource sizes. (1) It is possible that we do not reach a co-occurrence coverage of stimulus-associate pairs of 100% just because our corpus is not large enough. We cannot prove that this is not the case because there is always a size limit of a corpus. The best way to address this concern would be to use the currently largest available corpus, the World Wide Web, to check the co-occurrences of stimulus-associate pairs. However, we believe that even with the largest corpus one would probably not reach 100% coverage. This assumption was recently strengthened by Schulte im Walde and Melinger (to appear) who demonstrated that the co-occurrence of verb-association pairs (from the same source as in this paper) increases with an increasing corpus, but the increase becomes smaller with a larger corpus, thus it seems to reach a ceiling. (2) The corpus is much larger than the dictionary and Wikipedia articles, where the length of the articles varied between 4 and 3,177 words in WDG, and between 33 and 21,071 words in Wikipedia. In total, including all articles we downloaded for WDG and Wikipedia, the articles contained 593,318 words in WDG and 2,848,150 words in Wikipedia. Related to the differences in the sizes of the articles and the resource parts we worked with, of course the absolute frequencies of the associates we searched for (and therefore also the a priori probabilities to find an association by itself and within a specific stimulus entry) also varied strongly. It is difficult to overcome the frequency difficulties, though. One could adjust the corpus size to the sizes of the articles (i.e., only use a random part of the corpus whose size roughly corresponds to the sizes WDG and Wikipedia provide), but corpus co-occurrence is not expected to be as focused information as a dictionary or encyclopaedia article, so this would not do justice to the corpus analysis. Alternatively, one could provide statistical association scores (such as log-likelihood) for the stimulus-associate pairs within the resources. This kind of information would add a measure of how strong the associations between stimuli and responses are, according to their expected vs. observed frequencies within the resources. However, this information does not refer to our overall question within this paper, *whether* we find the stimulus-response pairs within a certain resource or not. In conclusion, we cannot easily overcome the size and frequency issues related to this article. As a temporary solution, we added the resource frequencies to all example stimuli and associates in the appendix. The frequencies illustrate that a stimulus-response pair that is missing within a certain resource is not necessarily missing because the resource frequency of the respective response is too low. For example, the association *Krieg* 'war' appears 136 times in the dictionary "corpus" (i.e., the sum of all articles we downloaded), but it is not found within the entry of the

stimulus *Soldat* 'soldier'; the association *Licht* 'light' appears 206 times in the Wikipedia "corpus", but not within the article of the stimulus *Lampe* 'lamp'; the stimulus *abspecken* 'lose weight' appears 200 times in the corpus, and the association *Gewicht* 'weight' appears 4,591 times in the corpus, but they are not found in co-occurrence with each other. Thus, there must be additional, resource-dependent reasons why a stimulus-associate pair is not covered by one of the resources.

5. Conclusion

This work presented an analysis of association norms that checked on how many and which of the semantic associates were found within corpus co-occurrence vs. dictionary and encyclopaedic articles of the respective stimulus words. We demonstrated that the resources differ in a) the parts-of-speech (nouns vs. verbs) they provide information for; b) the parts-of-speech (nouns vs. verbs vs. adjectives vs. adverbs) they predominantly use within their word descriptions, and c) the strength of the semantic relatedness between described and describing words (taking association strength as an indicator of semantic relatedness).

Even though we did not find evidence for our hypothesis that dictionary and encyclopaedic information provides more world knowledge than corpus co-occurrence, the information we found in the three resource types complements each other, to various degrees. Even taking the various sizes of the resources into account, our results indicate that a combination of the used types of knowledge resources should be helpful for distributional descriptions in data-intensive semantics, to empirically model word meaning and word similarity.

6. References

Bar-Hillel, Yehoshua (1960). The Present Status of Automatic Translation of Languages. *Advances in Computers*.

Engelberg, Stefan and Lothar Lemnitzer (2004). *Lexikographie und Wörterbuchbenutzung.* Stauffenburg-Verlag, Tübingen.

Gabrilovich, Evgeniy, and Shaul Markovitch (2006). Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence.*

Harris, Zellig (1968). Distributional Structure. In Jerold J. Katz (Eds.), *The Philosophy of Linguistics*, pp. 26-47. Oxford University Press.

Lesk, Michael (1986). Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the* SIGDOC *Conference.*

Lin, Dekang (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th Conference on Computational Linguistics.*

Lowe, Will, and Scott McDonald (2000). The Direct Route: Mediated Priming in Semantic Space. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.

Lund, Kevin, Curt Burgess, and Ruth Ann Atchley (1995). Semantic and Associative Priming in High-Dimensional Semantic Space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society of America*.

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

Mihalcea, Rada (2007). Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Padó, Sebastian, and Mirella Lapata (2007). Dependency-based Construction of Semantic Space Models. *Computational Linguistics* 33(2).

Poesio, Massimo, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira (2002). Acquiring Lexical Knowledge for Anaphora Resolution. In *Proceedings of the 3rd Conference on Language Resources and Evaluation.*

Ponzetto, Simone Paolo, and Michael Strube (2007). Knowledge Derived from Wikipedia For Computing Semantic Relatedness. *Journal of Artificial Intelligence Research* 30.

Sahlgren, Magnus (2006). *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces.* Ph.D. thesis. Stockholm University.

Salton, Gerard, Anita Wong, and Chung Sue Yang (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM* 18(11).

Schütze, Hinrich (1998). Automatic Word Sense Discrimination. *Computational Linguistics.* Special Issue on Word Sense Disambiguation.

Schulte im Walde, Sabine (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics* 32(2).

Schulte im Walde, Sabine, Alissa Melinger, Michael Roth, and Andrea Weber (to appear). An Empirical Characterisation of Response Types in German Association Norms. *Research on Language and Computation.*

Schulte im Walde, Sabine, and Alissa Melinger (to appear). An In-Depth Look into the Co-Occurrence Distribution of Semantic Associates. *Italian Journal of Linguistics. Special Issue on „From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science".*

Vigliocco, Gabriella, David Vinson, William Lewis, and Merrill Garrett (2004). Representing the Meanings of Object and Action Words: The Featural and Unitary Semantic Space Hypothesis. *Cognitive Psychology* 48.

Weber, Nico, editor (1996). *Semantik, Lexikographie und Computeranwendungen.* Max Niemeyer Verlag.

7. Appendix

This appendix lists the 10 strongest stimulus-associate pairs for each cross-comparison of resources. The pairs are accompanied by the association strength (i.e., how often an association was provided in response to a certain stimulus; column 1), the part-of-speech of the association (column 4), as well as the frequencies of the stimuli and the associations within the respective resources (in brackets following the stimulus/associate words).

7.1 **Associations to noun stimuli**

Wikipedia: yes – Corpus: no			
78	*Trauben* 'grapes' (184/723)	*Wein* 'wine' (78/4,246)	N
77	*Rutsche* 'slide' (2/0)	*Kind* 'child' (699/114,109)	N
76	*Wagon* 'carriage' (1/21)	*Zug* 'train' (616/19,750)	N
73	*Skier* 'skis' (1/0)	*Schnee* 'snow' (46/2,990)	N
67	*Rutsche* 'slide' (2/0)	*Spielplatz* 'playground' (5/2,764)	N
57	*Brezel* 'pretzel' (41/252)	*Salz* 'salt' (47/2,016)	N
48	*Zelt* 'tent' (34/3,154)	*Camping* 'camping' (4/68)	N
48	*Karotte* 'carrot' (25/165)	*orange* 'orange' (10/180)	ADJ
45	*Zebra* 'zebra' (19/161)	*Streifen* 'stripes' (67/137)	N
44	*Feuerwehrwagen* 'fire engine' (1/42)	*rot* 'red' (281/29,459)	ADJ

Dictionary: yes – Corpus: no			
48	*Schultafel* 'blackboard' (1/40)	*Lehrer* 'teacher' (65/13,359)	N
44	*Reibe* 'grater' (2/5)	*Käse* 'cheese' (99/1,137)	N
27	*Waschtisch* 'washstand' (1/25)	*waschen* 'wash' (46/2,191)	V
23	*Mütze* 'cap' (18/29)	*Wolle* 'wool' (15/398)	N
22	*Tiger* 'tiger' (4/1,329)	*Streifen* 'stripes' (5/137)	N
22	*Kegel* 'pin' (11/302)	*Bahn* 'alley' (48/11,863)	N
19	*Waschtisch* 'washstand' (1/25)	*Seife* 'soap' (10/544)	N
19	*Büroklammer* 'paper clip' (1/58)	*Büro* 'office' (11/11,922)	N
17	*Trichter* 'funnel' (4/159)	*einfüllen* 'fill in' (2/3,020)	V
17	*Tretroller* 'scooter' (1/13)	*fahren* 'drive' (560/36,188)	V

Corpus: yes – Wikipedia: no			
89	*Iglu* 'igloo' (58/16)	*Eskimo* 'Eskimo' (295/1)	N
86	*Filter* 'filter' (557/246)	*Kaffee* 'coffee' (5,349/251)	N
85	*Rüstung* 'armour' (993/45)	*Ritter* 'knight' (1,924/969)	N
81	*Lampe* 'lamp' (1,272/32)	*Licht* 'light' (13,806/206)	N
77	*Teleskop* 'telescope' (201/99)	*Stern* 'star' (7,243/302)	N
75	*Stecker* 'connector' (167/29)	*Strom* 'electricity' (6,937/123)	N
75	*Sandale* 'sandal' (163/35)	*Sommer* 'summer' (17,755/150)	N
74	*Wiege* 'cradle' (1,211/15)	*Baby* 'baby' (3,319/33)	N
74	*Schal* 'scarf' (552/16)	*Winter* 'winter' (8,966/167)	N
73	*Fackel* 'torch' (480/18)	*Feuer* 'fire' (10,017/172)	N

Corpus: yes – Dictionary: no			
95	*Schnuller* 'soother' (133/1)	*Baby* 'baby' (3319/10)	N
89	*Iglu* 'igloo' (58/1)	*Eskimo* 'Eskimo' (295/1)	N
88	*Schultafel* 'blackboard' (40/1)	*Kreide* 'chalk' (783/6)	N
85	*Zitrone* 'lemon' (418/4)	*sauer* 'sour' (2,695/16)	ADJ
85	*Kamel* 'camel' (515/10)	*Wüste* 'desert' (27,737/11)	N
79	*Thron* 'throne' (695/8)	*König* 'king'(11,037/47)	N
78	*Zitrone* 'lemon' (418/4)	*gelb* 'yellow' (5,716/26)	ADJ
77	*Teleskop* 'telescope' (201/1)	*Stern* 'star' (7,243/36)	N
77	*Stadion* 'stadium' (3675/6)	*Fußball* 'football' (16/10)	N
77	*Soldat* 'soldier' (28,043/37)	*Krieg* 'war' (42,466/136)	N

Wikipedia: yes – Dictionary: no			
95	*Schnuller* 'soother' (10/1)	*Baby* 'baby' (33/10)	N
88	*Schultafel* 'blackboard'	*Kreide* 'chalk' (26/6)	N
85	*Zitrone* 'lemon' (33/4)	*sauer* 'sour' (46/16)	ADJ
85	*Kamel* 'camel' (93/10)	*Wüste* 'desert' (103/11)	N
79	*Thron* 'throne' (71/8)	*König* 'king' (581/47)	N
78	*Zitrone* 'lemon' (33/4)	*gelb* 'yellow' (141/26)	ADJ
78	*Trauben* 'grapes' (184/7)	*Wein* 'wine' (78/95)	N
77	*Stadion* 'stadium' (77/6)	*Fußball* 'football' (76/10/)	N
77	*Soldat* 'soldier' (203/37)	*Krieg* 'war' (320/136)	N
77	*Rutsche* 'slide' (4/2)	*Kind* 'child' (699/419)	N

Dictionary: yes – Wikipedia: no			
86	*Filter* 'filter' (2/246)	*Kaffee* 'coffee' (60/251)	N
85	*Rüstung* 'armour' (19/45)	*Ritter* 'knight' (15/969)	N
81	*Lampe* 'lamp' (22/32)	*Licht* 'light' (80/206)	N
72	*Schnecke* 'snail' (10/138)	*langsam* 'slow' (94/201)	ADV
72	*Radio* 'radio' (14/162)	*Musik* 'music' (33/367)	N
72	*Dach* 'roof' (66/112)	*Ziegel* 'brick' (10/36)	N
70	*Straße* 'street' (137/391)	*Auto* 'car' (105/82)	N
69	*Topf* 'pot' (19/40)	*kochen* 'cook' (50/33)	V
64	*Kirsche* 'cherry' (19/23)	*rot* 'red' (85/281)	ADJ
62	*Zeitung* 'newspaper' (64/319)	*lesen* 'read' (97/143)	V

7.2 Associations to verb stimuli

Wikipedia: yes – Corpus: no			
21	*weinen* 'cry' (6/1,988)	*Tränen* 'tears' (13/0)	N
13	*einfrieren* 'freeze' (8/1,119)	*Kühlschrank* 'fridge' (21/1,755)	N
12	*fasten* 'fast' (10/170)	*Religion* 'religion' (224/4,055)	N
11	*trocknen* 'dry' (43/550)	*Trockner* 'dryer' (2/42)	N
11	*fasten* 'fast' (10/170)	*Diät* 'diet' (40/1,215)	N
10	*paddeln* 'paddle' (6/105)	*Paddel* 'paddle' (31/66)	N
8	*paddeln* 'paddle' (6/105)	*Ruder* 'oar' (53/881)	N
8	*heulen* 'cry' (4/836)	*Heulsuse* 'crybaby' (1/16)	N
8	*auftauen* 'defrost' (1/131)	*Wärme* 'heat' (87/1,610)	N
8	*abspecken* 'lose weight' (2/200)	*Gewicht* 'weight' (310/4,591)	N

Dictionary: yes – Corpus: no			
21	*weinen* 'cry' (27/1,988)	*Tränen* 'tears' (1/0)	N
20	*schwanen* 'bode ill' (1/6)	*ahnen* 'guess' (9/3,866)	V
12	*fasten* 'fast' (6/170)	*Religion* 'religion' (7/4,055)	N
11	*schwanen* 'guess' (1/6)	*Schwan* 'swan' (14/705)	N
11	*eilen* 'hurry' (31/2031)	*beeilen* 'hurry up' (11/754)	V
10	*paddeln* 'paddle' (6/105)	*Paddel* 'paddle' (1/66)	N
10	*aushaken* 'unhook' (1/0)	*einhaken* 'hook into' (11/49)	V
10	*aushaken* 'unhook' (1/0)	*Haken* 'hook' (26/1,290)	N
8	*grauen* 'dread' (18/214)	*Morgengrauen* 'dawn' (3/530)	N
8	*treiben* 'drove' (120/13,445)	*Herde* 'herd' (17/0)	N

Corpus: yes – Wikipedia: no			
47	*donnern* 'thunder' (792/3)	*Gewitter* 'thunderstorm' (1,037/56)	N
45	*gehen* 'walk' (22,7492/1,244)	*laufen* 'run' (37,860/280)	V
45	*fahren* 'drive' (36,188/243)	*Auto* 'car' (38,297/82)	N
45	*bauen* 'build' (27,853/478)	*Haus* 'house' (90,214/438)	N
43	*heulen* 'howl' (836/4)	*weinen* 'cry' (1,988/6)	V
42	*zahlen* 'pay' (22,910/49)	*Geld* 'money' (61,141/265)	N
42	*fliegen* 'fly' (10,483/127)	*Flugzeug* 'airplane' (8,073/292)	N
40	*schmelzen* 'melt' (692/21)	*Eis* 'ice' (4,142/107)	N
39	*kriegen* 'get' (8,394/3)	*bekommen* 'receive' (55,227/239)	V
36	*schleichen* 'sneak' (1,223/7)	*leise* 'silent' (1,934/13)	ADJ

Corpus: yes – Dictionary: no			
47	*donnern* 'thunder' (792/19)	*Gewitter* 'thunderstorm' (1,037/25)	N
42	*mampfen* 'munch' (176/1)	*essen* 'eat' (6,085/99)	V
39	*kriegen* 'get' (8,394/69)	*bekommen* 'receive' (55,227/142)	V
37	*lehren* 'teach' (3,875/18)	*Lehrer* 'teacher' (13,359/65)	N
37	*injizieren* 'inject' (163/3)	*Spritze* 'injection' (540/6)	N
36	*heilen* 'heal' (1,257/19)	*Arzt* 'doctor' (20,321/114)	N
34	*lehren* 'teach' (3,875/18)	*Schule* 'school' (21,739/58)	N
34	*erschrecken* 'scare' (1,597/46)	*Angst* 'fear' (28,330/61)	N
33	*unterrichten* 'instruct' (4,073/32)	*Schule* 'school' (21,739/58)	N
33	*schneien* 'snow' (367/18)	*kalt* 'cold' (10,986/68)	ADJ

Wikipedia: yes – Dictionary: no			
37	*lehren* 'teach' (71/18)	*Lehrer* 'teacher' (149 (1/65)	N
34	*lehren* 'teach' (71/18)	*Schule* 'school' (222/58)	N
31	*abspecken* 'lose weight' (2/0)	*Diät* 'diet' (40/5)	N
30	*abnehmen* 'lose weight' (53/49)	*Diät* 'diet' (40/5)	N
29	*atmen* 'breathe' (21/17)	*Lunge* 'lung' (45/8)	N
27	*lernen* 'study, learn' (173/72)	*Schule* 'school' (222/58)	N
25	*paddeln* 'paddle' (6/6)	*Wasser* 'water' (892/259)	N
25	*beginnen* 'begin' (693/127)	*Anfang* 'start' (463/19)	N
24	*rudern* 'row' (17/16)	*Wasser* 'water' (892/259)	N
23	*singen* 'sing' (85/49)	*Musik* 'music' (367/33)	N

Dictionary: yes – Wikipedia: no			
45	*gehen* 'walk' (1,253/1,244)	*laufen* 'run' (223/280)	V
45	*fahren* 'drive' (568/243)	*Auto* 'car' (105/82)	N
45	*bauen* 'build' (96/478)	*Haus* 'house' (383/438)	N
43	*heulen* 'howl' (23/4)	*weinen* 'cry' (27/6)	V
42	*zahlen* 'pay' (42/49)	*Geld* 'money' (323/265)	N
42	*fliegen* 'fly' (135/127)	*Flugzeug* 'airplane' (113/292)	N
40	*schmelzen* 'melt' (34/21)	*Eis* 'ice' (62/107)	N
36	*schleichen* 'sneak' (15/7)	*leise* 'quiet' (20/7)	ADJ
36	*kosten* 'cost' (105/45)	*Geld* 'money' (323/265)	N
36	*frieren* 'freeze' (68/3)	*kalt* 'cold' (68/128)	ADJ