A Database of Paradigmatic Semantic Relation Pairs for German Nouns, Verbs, and Adjectives

Silke Scheible and Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
{scheible,schulte}@ims.uni-stuttgart.de

Abstract

A new collection of semantically related word pairs in German is presented, which was compiled via human judgement experiments and comprises (i) a representative selection of target lexical units balanced for semantic category, polysemy, and corpus frequency, (ii) a set of human-generated semantically related word pairs based on the target units, and (iii) a subset of the generated word pairs rated for their relation strength, including positive and negative relation evidence. We address the three paradigmatic relations *antonymy, hypernymy* and *synonymy*, and systematically work across the three word classes of adjectives, nouns, and verbs.
A series of quantitative and qualitative analyses demonstrates that (i) antonyms are more canonical than hypernyms and synonyms, (ii) relations are more or less natural with regard to the specific word classes, (iii) antonymy is clearly distinguishable from hypernymy and synonymy, but hypernymy and synonymy are often confused. We anticipate that our new collection of semantic relation pairs will not only be of considerable use in computational areas in which semantic relations play a role, but also in studies in theoretical linguistics and psycholinguistics.

1 Introduction

This paper describes the collection of a database of paradigmatically related word pairs in German which was compiled via human judgement experiments hosted on Amazon Mechanical Turk. While paradigmatic relations (such as synonymy, antonymy, hypernymy, and hyponymy) have been extensively researched in theoretical linguistics and psycholinguistics, they are still notoriously difficult to identify and distinguish computationally, because their distributions in text tend to be very similar. For example, in *The boy/girl/person loves/hates the cat*, the nominal co-hyponyms *boy, girl* and their hypernym *person* as well as the verbal antonyms *love* and *hate* occur in identical contexts, respectively. A dataset of paradigmatic relation pairs would thus represent a valuable test-bed for research on semantic relatedness.

For the compilation of the relation dataset we aimed for a sufficiently large amount of human-labelled data, which may both serve as seeds for a computational approach, and provide a gold-standard for evaluating the resulting computational models. This paper describes our efforts to create such a ***paradigmatic relation dataset in a two-step process***, making use of two types of human-generated data: (1) human suggestions of semantically related word pairs, and (2) human ratings of semantic relations between word pairs. Furthermore, we are the first to ***explicitly work across word classes (covering adjective, noun and verb targets)***, and to ***incorporate semantic classes, corpus frequency and polysemy as balancing criteria into target selection***. The resulting dataset[1] consists of three parts:

1. A representative selection of target lexical units drawn from GermaNet, a broad-coverage lexical-semantic net for German, using a principled sampling technique and taking into account the three major word classes adjectives, nouns, and verbs, which are balanced according to semantic category, polysemy, and type frequency.
2. A set of human-generated semantically related word pairs, based on the target lexical units.
3. A subset of semantically related word pairs, rated for the strength of the relation between them.

[1]The dataset is available from http://www.ims.uni-stuttgart.de/data/sem-rel-database.

We anticipate that our new collection of semantic relation pairs will not only be of considerable use in computational areas in which semantic relations play a role (such as Distributional Semantics, Natural Language Understanding/Generation, and Opinion Mining), but also in studies in theoretical linguistics and psycholinguistics. In addition, our dataset may be of major interest for research groups working on automatic measures of semantic relatedness, as it allows a principled evaluation of such tools. Finally, since the target lexical units are drawn from the GermaNet database, our results will be directly relevant for assessing, developing, and maintaining this resource.

2 Related work

Over the years a number of datasets have been made available for studying and evaluating semantic relatedness. For English, Rubenstein and Goodenough (1965) obtained similarity judgements from 51 subjects on 65 noun pairs, a seminal study which was later replicated by Miller and Charles (1991), and Resnik (1995). Finkelstein et al. (2002) created a set of 353 English noun-noun pairs rated by 16 subjects according to their semantic relatedness on a scale from 0 to 10. For German, Gurevych (2005) replicated Rubenstein and Goodenough's experiments by translating the original 65 word pairs into German. In later work, she used the same experimental setup to increase the number of word pairs to 350 (Gurevych, 2006).

The dataset most similar to ours is *BLESS* (Baroni and Lenci, 2011), a freely available dataset that includes 200 distinct English concrete nouns as target concepts, equally divided between living and non-living entities, and grouped into 17 broad classes such as *bird, fruit*. For each target concept, BLESS contains several relata, connected to it through a semantic relation (hypernymy, co-hyponymy, meronymy, attribute, event), or through a null-relation. BLESS thus includes two paradigmatic relations (hypernymy, co-hyponymy) but does not focus on paradigmatic relations. Furthermore, it is restricted to concrete nouns, rather than working across word classes.

3 Paradigmatic relations

The focus of this work is on semantic relatedness, and in particular on paradigmatic semantic relations. This section discusses the theoretical background of the notion of *paradigmatic semantic relations*. The term **paradigmatic** goes back to de Saussure (1916), who introduced a distinction between linguistic elements based on their position relative to each other. This distinction derives from the linear nature of linguistic elements, which is reflected in the fact that speech sounds follow each other in time. Saussure refers to successive linguistic elements that combine with each other as 'syntagma', and thus the relation between these elements is called 'syntagmatic'. On the other hand, elements that can be found in the same position in a syntagma, and which could be substituted for each other, are in a 'paradigmatic' relationship. While syntagmatic and paradigmatic relations can hold between a variety of linguistic units (such as morphemes, phonemes, clauses, or sentences), the focus of this research is on the relations between words.

Many studies in computational linguistics work on the assumption that paradigmatic semantic relations hold between words. As will become apparent in the course of this work, it is necessary to move beyond these definitions for an appropriate investigation of paradigmatic semantic relations. According to Cruse (1986), sense is defined as "the meaning aspect of a lexical unit", and he states that "semantic relations" hold between lexical units, not between lexemes.

The goal of this work is to create a database of semantic relations for German adjectives, nouns and verbs, focussing on the three types of paradigmatic relations referred to as *sense-relations* by Lyons (1968, 1977): synonymy, antonymy, and hypernymy.

4 Experimental setup

Our aim was to collect semantically related word pairs for the paradigmatic relations antonymy, synonymy, and hypernymy for the three word classes nouns, verbs, and adjectives. For this purpose we implemented two experiments involving human participants. Starting with a set of target words, in the first experiment participants were asked to propose suitable antonyms, synonyms and hypernyms for

each of the targets. For example, for the target verb *befehlen* ('to command'), participants proposed antonyms such as *gehorchen* ('to obey'), synonyms such as *anordnen* ('to order'), and hypernyms such as *sagen* ('to say').

In the second experiment, participants were asked to rate the strength of a given semantic relation with respect to a word pair on a 6-point scale. For example, workers would be presented with a pair "*wild – free*" and asked to rate the strength of antonymy between the two words. All word pairs were assessed with respect to all three relation types.

Both experiments will be described in further detail in Sections 5 and 6. The current section aims to provide an overview of GermaNet, a lexical-semantic word net for German, from which the set of target words was drawn (4.1). We then describe the selection of target words from GermaNet, which used a stratified sampling approach (4.2). Finally, we introduce the platform used to implement the experiments, Amazon Mechanical Turk (4.3).

4.1 Target source: GermaNet

GermaNet is a lexical-semantic word net that aims to relate German nouns, verbs, and adjectives semantically. GermaNet has been modelled along the lines of the Princeton WordNet for English (Miller et al., 1990; Fellbaum, 1998) and shares its general design principles (Hamp and Feldweg, 1997; Kunze and Wagner, 1999; Lemnitzer and Kunze, 2007). For example, lexical units denoting the same concept are grouped into synonym sets ('synsets'). These are in turn interlinked via conceptual-semantic relations (such as hypernymy) and lexical relations (such as antonymy). For each of the major word classes, the databases further take a number of semantic categories into consideration, expressed via top-level nodes in the semantic network (such as 'Artefakt/artifact', 'Geschehen/event', 'Gefühl/feeling'). However, in contrast to WordNet, GermaNet also includes so-called 'artificial concepts' to fill lexical gaps and thus enhance network connectivity, and to avoid unsuitable co-hyponymy (e.g. by providing missing hypernyms or hyponyms). GermaNet also differs from WordNet in the way in which it handles part of speech. For example, while WordNet employs a clustering approach to structuring adjectives, GermaNet uses a hierarchical structure similar to the one employed for the noun and verb hierarchies. Finally, the latest releases of WordNet and GermaNet also differ in size: While WordNet 3.0 contains at total of 117,659 synsets and 155,287 lexical units, the respective numbers for GermaNet 6.0 are considerably lower, with 69,594 synsets and 93,407 lexical units.

Since GermaNet is the largest database of its kind for German, and as it encodes all types of relations that are of interest for us (synonymy, antonymy, and hypernymy), it represents a suitable starting point for our purposes.

4.2 Target selection

The purpose of collecting the set of targets was to acquire a broad range of lexical items which could be used as input for generating semantically related word pairs (cf. Section 5). Relying on GermaNet version 6.0 and the respective *JAVA API*, we used a stratified sampling technique to randomly select 99 nouns, 99 adjectives and 99 verbs from the GermaNet files. The random selection was balanced for

1. the ***size of the semantic classes***,[2] accounting for the 16 semantic adjective classes and the 23 semantic classes for both nouns and verbs, as represented by the file organisation;
2. ***three polysemy classes*** according to the number of GermaNet senses:
   I) monosemous, II) two senses and III) more than two senses;
3. ***three frequency classes*** according to type frequency in the German web corpus *SdeWaC* (Faaß and Eckart, 2013), which contains approx. 880 million words:
   I) *low* (200–2,999), II) *mid* (3,000–9,999) and III) *high* ($\geq$10,000).

The total number of 99 targets per word class resulted from distinguishing 3 sense classes and 3 frequency classes, $3 \times 3 = 9$ categories, and selecting 11 instances from each category, in proportion to the semantic class sizes.

[2] For example, if an adjective GermaNet class contained 996 word types, and the total number of adjectives over all semantic classes was 8,582, and with 99 stimuli collected in total, we randomly selected $99 * 996/8,582 = 11$ adjectives from this class.

4.3 Experimental platform: Mechanical Turk

The experiments described below were implemented in Amazon Mechanical Turk (AMT)[3], a web-based crowdsourcing platform which allows simple tasks (so-called HITs) to be performed by a large number of people in return for a small payment. In our first experiment, human associations were collected for different semantic relation types, where AMT workers were asked to propose suitable synonyms, antonyms, and hypernyms for each of the targets. The second experiment was based on a subset of the generated synonym/antonym/hypernym pairs and asked the workers to rate each pair for the strength of antonymy, synonymy, and hypernymy between them, on a scale between 1 (minimum strength) and 6 (maximum strength). To control for non-native speakers of German and spammers, each batch of HITs included two examples of 'non-words' (invented words following German morphotactics such as *Blapselheit*, *gekortiert*) in a random position. If participants did not recognise the invented words, we excluded all their ratings from consideration. While we encouraged workers to complete all HITs in a given batch, we also accepted a smaller number of submitted HITs, as long as the workers had a good overall feedback score.

5 Generation experiment

5.1 Method

The goal of the generation experiment was to collect human associations for the semantic relation types antonymy, hypernymy, and synonymy. For each of our 3×99 adjective, noun, and verb targets, we asked 10 participants to propose a suitable synonym, antonym, and hypernym. Targets were bundled randomly in 9 batches per word class, each including 9 targets plus two invented words. The experiment consisted of separate runs for each relation type to avoid confusion between them, with participants first generating synonyms, then antonyms, and finally hypernyms for the targets, resulting in 3 word classes $\times$ 99 targets $\times$ 3 relations $\times$ 10 participants $= 8,910$ target–response pairs.

5.2 Results and discussion

5.2.1 Total number of responses

Table 1 illustrates how the number of generated word pairs distributes across word classes and relations. The total number per class and relation is 990 tokens (99 targets $\times$ 10 participants). From the maximum number of generated pairs, a total of 131 types (211 tokens) were discarded because the participants provided no response. These cases had been accepted via AMT nevertheless because the participants were approved workers and we assumed that the empty responses showed the difficulty of specific word–relation constellations, cf. Section 5.2.3. For example, six out of ten participants failed to provide a synonym for the adjective *bundesrepublikanisch* 'federal republic'.

	ANT		HYP		SYN		*all*	
	types	tokens	types	tokens	types	tokens	types	tokens
ADJ	524	990	676	990	597	990	1,797	2,970
NOUN	708	990	701	990	621	990	2,030	2,970
VERB	636	990	662	990	620	990	1,918	2,970
all	1,868	2,970	2,039	2,970	1,838	2,970	5,745	8,910

Table 1: Number of generated relation pairs across word classes.

5.2.2 Number of ambiguous responses

An interesting case is provided by pairs that were generated with regard to different relations but for the same target word. Table 2 lists the number of types of such ambiguous pairs, and the intersection of the tokens. For example, if five participants generated a pair with regard to a target and relation x, and two participants generated the same pair with regard to relation y, the intersection is 2. The intersection

[3] https://www.mturk.com

is more indicative of ambiguity here, because in most cases of ambiguity the intersection is only 1, which might as well be the result of an erroneously generated pair (e.g., because the participant did not pay attention to the task), rather than genuine ambiguity. Examples of ambiguous responses with an intersection > 1 are *Gegenargument–Argument* 'counter argument – argument', which was provided five times as an antonymy pair and twice as a hypernymy pair; *freudlos–traurig* 'joyless – sad', which was provided four times as a synonymy pair and five times as a hypernymy pair; and *beseitigen–entfernen* 'eliminate – remove', which was provided five times as a synonymy pair and five times as a hypernymy pair.

	ANT+HYP		ANT+SYN		HYP+SYN		ANT+HYP+SYN	
	types	tokens	types	tokens	types	tokens	types	tokens
ADJ	6	6	4	4	195	342	2	2
NOUN	15	16	17	17	93	117	5	6
VERB	4	4	8	8	182	290	5	6
all	25	26	29	29	470	749	12	14

Table 2: Number of ambiguous relation pairs across word classes.

The ambiguities in Table 2 indicate that humans are quite clear about what distinguishes antonyms from synonyms, and what distinguishes antonyms from hypernyms. On the other hand, the line dividing hypernymy and synonymy is less clear, and the large amount of confusion between the two relations lends support to theories claiming that hypernymy should be considered a type of synonymy, and that real synonymy does not exist in natural languages for economical reasons. Furthermore, the confusion is most obvious for adjectives and verbs, for which the relation is considered less natural than for nouns, cf. Miller and Fellbaum (1991).

5.2.3 Number of (different) responses across word classes and relations

An analysis of the number of different antonyms, hypernyms and synonyms generated for a given target shows no noticeable difference at first glance: on average, 6.04 different antonyms were generated for the targets, while the number is minimally higher for synonyms with 6.08 different responses on average; hypernyms received considerably more (6.78) different responses on average. However, the distribution of the numbers of different antonym, hypernym, and synonym responses across the targets shows that the antonymy generation task results in more targets with a small number of different responses compared to the synonymy and the hypernym task (Figure 1): there are 10 targets for which all ten participants generated the same antonym (x = *number of different responses* = 1), such as *dunkel–hell* 'dark – light' and *verbieten–erlauben* 'to forbid – to allow', while there are 17 targets where they generated exactly two (x=2), and 29 targets where they suggested three different antonyms (x=3). In contrast, for hypernymy and synonymy, there are 0/3 targets where all participants agreed on the same response, and there are only 5/10 targets where they generated exactly two, and 8/21 targets where they generated only three different hypernyms/synonyms.

These results are in line with previous findings for English and Swedish by Paradis and Willners (2006) and Paradis et al. (2009), who argue against the strict contrast between 'direct' and 'indirect' antonyms which has been assumed in the literature (see, for example, Gross et al. (1989)) in favour of a scale of 'canonicity' where some word pairs are perceived as more antonymic than others. In particular, they propose that the weaker the degree of canonicity, the more different responses the target items will yield in an elicitation experiment. Similar to the current findings for German, they found that for English and Swedish there is a small core of highly opposable couplings which have been conventionalised as antonym pairs in text and discourse, while all other couplings form a scale from more to less strongly related. The ten targets for which all participants generated the same antonym response are thus likely to represent highly "canonical" pairings. The fact that the hypernymy and synonymy generation experiments results in fewer targets with only one or two different responses suggests that hypernymy and synonymy have a lower level of canonicity than antonymy.

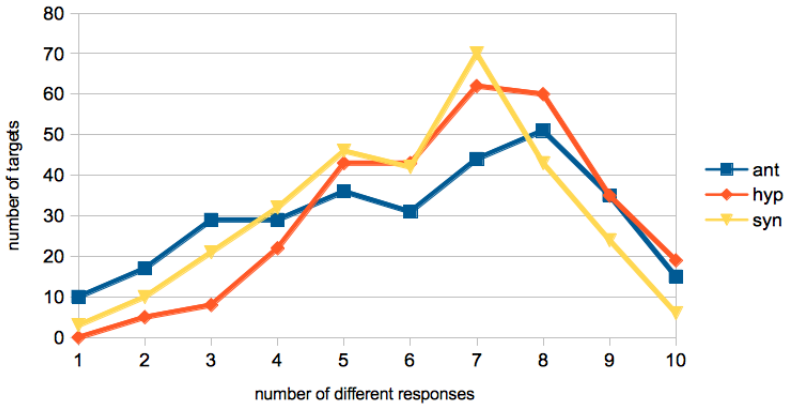

Figure 1: Number of targets plotted against the number of different responses.

Figure 2 demonstrates that the overall distributions of the frequency of responses are, however, very similar for antonyms, hypernyms and synonyms: between 72% and 77% of the responses were only given once, with the curves following a clear downward trend. Note that a strength of 10 in Figure 2 refers to the case of *one different response* (x=1) in Figure 1.

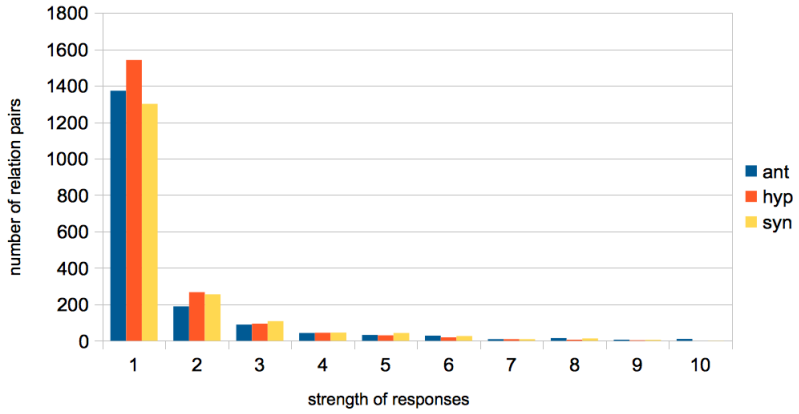

Figure 2: Response magnitude.

Finally, Figure 3 compares the number of blank responses (types and tokens) regarding antonyms, hypernyms and synonyms. Across word classes, 74/115 targets (types/tokens) received blank antonym responses, while only 25/34 targets received blank hypernym responses and only 32/62 targets received blank synonym responses. These numbers indicate that participants find it harder to come up with antonyms than hypernyms or synonyms. Breaking the proportions down by word class, Figure 3 demonstrates that in each case the number of missing antonyms (left panel: types; right panel: tokens) is larger than those of missing hypernyms/synonyms. Figure 3 also shows that the difficulty to provide relation pairs varies across word classes. While antonyms are the most difficult relation in general, there are more blank responses regarding adjectives and nouns, in comparison to verbs. Hypernymy seems similarly difficult across classes, and synonymy is more difficult for nouns than for adjectives or verbs.

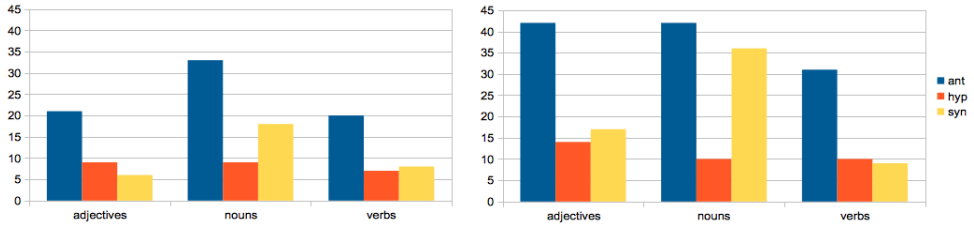

Figure 3: Blank responses (types and tokens).

5.2.4 Comparison with GermaNet

The results of the generation experiment can be used to extend and develop GermaNet, the resource the targets were drawn from: a large proportion of responses are not covered in GermaNet. Table 3 below shows for the three parts of speech and the three relation types how many responses were covered by both the generation experiment (EXP) and GermaNet (GN) (column 'Both'), how many of them only appear in the generation experiment (column 'EXP'), and how many are only listed in GermaNet ('GN'). Blank and multi-word responses in the experimental results were excluded from consideration. The comparison shows that the variety of semantic relation types in our experimental dataset is considerably larger than in GermaNet, while the overlap is marginal. Especially for antonyms, the coverage in GermaNet seems to be quite low, across word classes. For hypernymy and synonymy, the semantic relation types complement each other to a large extent, with each resource containing relations that are not part of the other resource. In sum, the tables confirm that extending GermaNet with our relation types should enrich the manual resource.

	ANT			HYP			SYN		
	Both	EXP	GN	Both	EXP	GN	Both	EXP	GN
ADJ	33	453	5	100	561	237	66	496	160
NOUN	3	633	2	108	561	393	59	516	150
VERB	10	542	2	132	507	260	40	554	109

Table 3: Relation coverage in Generation Experiment (EXP) and GermaNet (GN).

6 Rating experiment

6.1 Method

In the second experiment, Mechanical Turk workers were asked to rate the strength of a given semantic relation with respect to a word pair on a 6-point scale. The main purpose of this experiment was to identify and distinguish between "strong" and "weak" examples of a specific relation. The number of times a specific response was given in the generation experiment does not necessarily indicate the strength of the relation. This is especially true for responses that were suggested by only one or two participants, where it is difficult to tell if the response is an error, or if it relates to a idiosyncratic sense of the target word that the other participants did not think of in the first instance. Crucially, in the rating experiment all word pairs were assessed with respect to all three relation types, thus asking not only for positive but also negative evidence of semantic relation instances.

The set of word pairs used as input is a carefully selected subset of responses acquired in the generation experiment.[4] For each of the 99 targets and each of the semantic relations antonymy, synonymy, and hypernymy two responses were included (if available): the *response with the highest frequency* (random choice if several available), and a *response with a lower frequency* (2, if available, otherwise 1; random choice if several available). Multi-word responses and blanks were excluded from consideration. A manual post-processing step aimed to address the issue of duplicate pairs in the randomly generated dataset, where the same responses had been generated for two of the relations.

In theory, each target should have 6 associated pairs (2xANT, 2xHYP, 2xSYN). In practice, there are sometimes fewer than 6 pairs per target in the dataset, because (i) for some targets, only one response is available for a given relation (e.g., if all 10 participants provided the same response), or (ii) no valid response of the required frequency type is available. The resulting dataset includes 1,684 target-response pairs altogether, 546 of which are adjective pairs, 574 noun pairs, and 564 verb pairs. To avoid confusion, the ratings were collected in separate experimental settings, i.e., for each word class and each relation type, all generated pairs were first judged for their strength of one relation, and then for their strength of another relation.

[4] For time and money reasons, we could not collect the $8,910 \times 3 \times 10 = 267,300$ ratings for all responses.

6.2 Results and discussion

In the following, we present the results of the rating experiment in terms of mean rating scores for each word pair. The mean rating scores were calculated across all ten ratings per pair. The purpose of the analysis was to verify that the responses generated in the generation experiment are in fact perceived as examples of the given relation type by other raters. We thus looked at all responses for a given relation type in the data set and calculated the average value of all mean ratings for this relation type. For example, Figure 4 (left panel) shows that the responses generated as antonyms are clearly perceived as antonyms in the case of adjectives, with an average rating score of 4.95. Verb antonyms are also identified as such with a rating of 4.38. The situation for nouns, however, is less clear: an average rating of 3.70 is only minimally higher than the middle point of the rating scale (3.50). These findings support the common assumption that antonymy is a relation that applies well to adjectives and verbs, but less so to nouns. Responses generated as synonyms (plot omitted for space reasons), on the other hand, are identified as such for all three words classes, with average rating values of 4.78 for adjectives, 4.48 for nouns, and 4.66 for verbs.

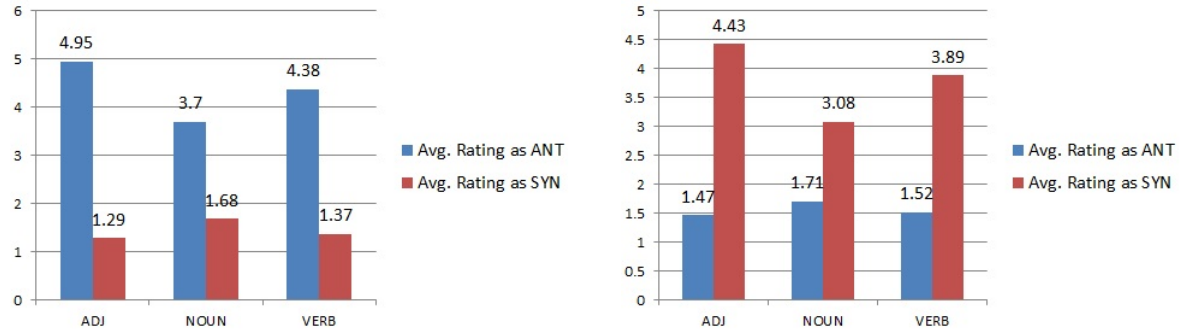

Figure 4: Average ratings of antonym/hypernym responses as ANT or SYN, across word classes.

Finally, Figure 4 (right panel) shows the average ratings as synonyms/antonyms for responses generated as hypernyms. The findings corroborate our analysis of synonym/hypernym confusion in Section 5: the distribution looks fairly similar to the one for synonyms, with low antonymy ratings, but an average synonymy rating of 4.43 for adjectives, 3.08 for nouns, and 3.89 for verbs. The results suggest that hypernymy is particularly difficult to distinguish from synonymy in the case of adjectives.

7 Conclusion

This article presented a new collection of semantically related word pairs in German which was compiled via human judgement experiments. The database consists of three parts:

1. A representative selection of target lexical units drawn from GermaNet, using a principled sampling technique and taking into account the three major word classes adjectives, nouns, and verbs, which are balanced according to semantic category, polysemy, and type frequency.
2. A set of 8,910 human-generated semantically related word pairs, based on the target lexical units.
3. A subset of 1,684 semantically related word pairs, rated for the strengths of relations.

To our knowledge, our dataset is the first that (i) focuses on multiple paradigmatic relations, (ii) systematically works across word classes, (iii) explicitly balances the targets according to semantic category, polysemy and type frequency, and (iv) explicitly provides positive and negative rating evidence. We described the generation and the rating experiments, and presented a series of quantitative and qualitative analyses. The analyses showed that (i) antonyms are more canonical than hypernyms and synonyms, (ii) relations are more or less natural with regard to the specific word classes, (iii) antonymy is clearly distinguishable from hypernymy and synonymy, and (iv) hypernymy and synonymy are often confused.

Acknowledgements

The research was funded by the DFG Sachbeihilfe SCHU-2580/2-1 (Silke Scheible) and the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde).

References

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed Distributional Semantic Evaluation. In *Proceedings of the EMNLP Workshop on Geometrical Models for Natural Language Semantics*, pages 1–10, Edinburgh, UK.

D. Allan Cruse. 1986. *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, UK.

Ferdinand de Saussure. 1916. *Cours de Linguistique Générale*. Payot.

Gertrud Faaß and Kerstin Eckart. 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.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Derek Gross, Ute Fischer, and George A. Miller. 1989. Antonymy and the Representation of Adjectival Meanings. *Memory and Language*, 28(1):92–106.

Iryna Gurevych. 2005. Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pages 767–778, Jeju Island, Korea.

Iryna Gurevych. 2006. Thinking beyond the Nouns - Computing Semantic Relatedness across Parts of Speech. In *Sprachdokumentation & Sprachbeschreibung, 28. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, Bielefeld, Germany.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Claudia Kunze and Andreas Wagner. 1999. Integrating GermaNet into EuroWordNet, a Multilingual Lexical-Semantic Database. *Sprache und Datenverarbeitung*, 23(2):5–19.

Lothar Lemnitzer and Claudia Kunze. 2007. *Computerlexikographie*. Gunter Narr Verlag, Tübingen, Germany.

George A. Miller and Walter G. Charles. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1):1–28.

George A. Miller and Christiane Fellbaum. 1991. Semantic Networks of English. *Cognition*, 41:197–229.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Carita Paradis and Caroline Willners. 2006. Antonymy and Negation: The Boundedness Hypothesis. *Journal of Pragmatics*, 38:1051–1080.

Carita Paradis, Caroline Willners, and Steven Jones. 2009. Good and Bad Opposites: Using Textual and Experimental Techniques to Measure Antonym Canonicity. *The Mental Lexicon*, 4(3):380–429.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, San Francisco, CA.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.