Designing a Database of GermaNet-based Semantic Relation Pairs involving Coherent Mini-Networks

Silke Scheible and Sabine Schulte im Walde

Institute for Natural Language Processing (IMS)
University of Stuttgart, Germany
scheible@ims.uni-stuttgart.de, schulte@ims.uni-stuttgart.de

Abstract

We describe the design and compilation of a new database containing German semantic relation pairs drawn from the lexical network GermaNet. The database consists of two parts: A representative selection of lexical units drawn from the three major word classes adjectives, nouns, and verbs, which are balanced according to semantic category, polysemy, and type frequency ('SemrelTargets'); and a set of semantically coherent GermaNet subnets consisting of semantic relations pairs clustering around the selected targets ('SemrelNets'). The database, which contains 99 SemrelTargets for each of the three word classes, and a total of 1623 relation pairs distributed across the respective subnets, promises to be an important resource not only for research in computational linguistics, but also for studies in theoretical linguistics and psycholinguistics. Currently, the data is being used in two types of human judgement experiments, one focusing on the generation of semantically related word pairs, and the other on rating the strength of semantic relations.

1. Introduction

Paradigmatic semantic relations such as synonymy, antonymy, hypernymy/hyponymy, and co-hyponymy have been the focus of many studies in theoretical and applied linguistics (Cruse (1986); Lyons (1977); Murphy (2003)). Approaches in computational linguistics also addressed paradigmatic relations, especially synonymy (e.g., Edmonds and Hirst (2002); Curran (2003); Lin et al. (2001)) and hypernymy (e.g., Hearst (1992); Caraballo (2001); Snow et al. (2004)), but less so antonymy, and often with respect to modelling contradiction (e.g., Lucerto et al. (2004); Harabagiu et al. (2006); de Marneffe et al. (2008)). Many approaches included one or the other paradigmatic relation within a set of target relations (e.g., Pantel and Pennacchiotti (2006); Morris and Hirst (2004); Turney (2006)), but to our knowledge no earlier work has specifically focused on all standard paradigmatic relations.

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). In 2001, 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).

Zesch and Gurevych (2006) note a number of shortcomings of previous approaches to creating datasets of semantic relatedness. First of all, they state that manually compiled lists of word pairs are often biased towards highly related pairs. They further draw attention to the fact that previous studies considered semantic relatedness of *words* rather than *concepts*, noting that polysemous or homonymous words should be annotated on the level of concepts. To overcome these limits for German, they propose automatic corpus-based methods which they employ to create a set of 328 related concept pairs across different word classes and drawn from three different domain-specific corpora. While this approach enables fast development of a large domain-specific dataset covering all types of lexical and semantic relations, they found that highly related concept pairs were under-represented in their data.

In this paper we describe the design and compilation of a new large-scale dataset containing German concept pairs related via paradigmatic semantic relations, which is currently being annotated with human judgements on the relations. Like Zesch and Gurevych (2006), our approach involves automatic compilation methods and a focus on concepts rather than words. However, in contrast to their approach, our data is drawn from GermaNet (Lemnitzer and Kunze, 2007), a broad-coverage lexical-semantic net for German, using a principled sampling technique. The resulting dataset consists of two parts:

1. A representative selection of lexical units drawn from the three major word classes adjectives, nouns, and verbs, which are balanced according to semantic category, polysemy, and type frequency (referred to as 'SemrelTargets'); and

2. A set of salient semantic GermaNet subnets consisting of paradigmatic semantic relations clustering around each of these targets ('SemrelNets').

The semantically coherent subnets (illustrated in Figure 1) allow an assessment of concepts within their semantic neighbourhood, and the stratified sampling technique ensures that the dataset contains a broad variety of relation pairs. The data is currently being used in two types of human judgement experiments: One focusing on the generation of semantically related word pairs, and the other on human rating of the strength of semantic relations.

The dataset contains a set of target lexical units (99 SemrelTargets each for the three word classes) and 1623 relation

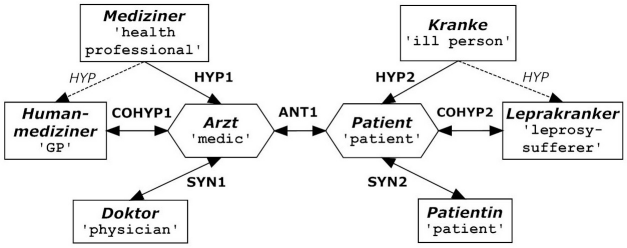

Figure 1: Example of a SemrelNet (Target *Arzt*, 'doctor')

pairs distributed across the respective subnets, thus representing one of the largest principled datasets for studying semantic relations. We anticipate that it 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.

This paper introduces the selection criteria and tools which were implemented to extract the set of SemrelTargets and their associated SemrelNets from GermaNet[1]. Section 2 aims to provide further motivation for the creation of this dataset by giving a brief overview of the research project it is part of, and discussing potential applications of this work. After introducing the database GermaNet, from which our data was sampled (Section 3), we describe the sampling procedure employed to select the set of target lexical units (Section 4). Section 5 deals with the notion of 'SemrelNets', and provides a detailed overview of the algorithm and associated tool for building these networks. Finally, in Section 6 we outline two human judgement experiments that are currently in progress, which are based on the dataset described in this paper.

2. Motivation

The compilation of the semantic relations dataset is part of a larger research project in the area of distributional semantics. One major goal of this project is to enhance computational work on paradigmatic semantic relations such as synonymy, antonymy, hypernymy, hyponymy, and co-hyponymy. While paradigmatic relations 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 sentence 'The boy/girl/person loves/hates his cat', the co-hyponyms *boy*, *girl*, and *person* as well as the antonyms *love* and *hate* occur in identical contexts. We are particularly interested in a theoretically and cognitively adequate selection of features to model word relatedness, paying special attention to word senses and any resulting ambiguities, an issue which is a well-known problem in computational linguistics in general, but which has been largely disregarded in distributional semantics.

[1]Both data and tools will be made freely available on our project homepage (http://www.ims.uni-stuttgart.de/projekte/semrel/resources.html).

In order to address these goals we require 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. In particular, we plan to make 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. The dataset described in this paper has been designed to enable these studies, and Section 6 will provide further details on the human judgement experiments carried out on the basis of this data.

While the dataset was designed with specific goals in mind, its general design and the associated extraction tools will also be of interest for other areas of NLP and linguistic research, for example Opinion Mining and Sentiment Analysis (where it is important to be aware of synonymy/hypernymy vs. antonymy in order to keep track of continuing vs. changing opinions/sentiments); Statistical Machine Translation (where it is important to be aware of the semantic relations between words because this can help in translation); and Word Sense Disambiguation (where the networks should be able to help with sense definitions in the Gold Standards). In addition, our dataset will also 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 our data is drawn from the GermaNet database, our results will be directly relevant for assessing, developing, and maintaining this resource. The random selection of SemrelTargets balanced by semantic category, number of senses and corpus frequency allows a systematic assessment of any biases in the semantic taxonomy. Coupled with further analyses, the evaluation can be as deep as the developer wants it to be. For example, we are currently analysing the random choices with respect to morphological properties, such as simplex vs. complex, and more specifically the types of noun compounds and particle verbs, etc. In the same vein, the SemrelNets point the developer to semantic areas that are particularly (non-) dense. Differences between densities in the networks are expected, they have been shown to be problematic in lexical hierarchies of this kind (Jiang and Conrath, 1997). The SemrelNets allow developers to systematically check if a very low/strong density is appropriate for a specific subnetwork, or if the network is under-/over-represented at that point.

3. GermaNet

GermaNet is a lexical-semantic word net that aims to relate German nouns, verbs, and adjectives semantically. GermaNet has been modelled on Princeton WordNet for English (Miller et al. (1990); Fellbaum (1998)) and shares its general design principles (Kunze and Wagner (1999); Lemnitzer and Kunze (2007)). For example, lexical units denoting the same concept are grouped into synonym sets (so-called '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

Senses	Freq	Gefühl	Verhalten
1	low	-	*satanisch*, 'satanic'; *gesprächsbereit* , 'ready to talk'
	mid	*empört*, 'indignant'	*naiv*, 'naive'; *schützend*, 'protective'
	high	*witzig*, 'funny'	*rassistisch*, 'racist'; *geschickt*, 'adept'
2	low	-	*drollig*, 'cute'
	mid	*reichhaltig*, 'rich'	*unruhig*, 'unsettled'
	high	*düster*, 'gloomy'	*unschuldig*, 'innocent'
3	low	*furios*, 'furious'	*erledigt*, 'done'
	mid	*heiter*, 'cheerful'	*faul*, 'lazy'; *energisch*, 'energetic'
	high	*wild*, 'wild'	*locker*, 'casual'; *mild*, 'mild'

Table 1: Selection of adjectival SemrelTargets for the semantic categories "Gefühl" ('feeling') and "Verhalten" ('behaviour') in GermaNet

nodes in the semantic network (such as 'Artefakt/artifact', 'Geschehen/event', or '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-hyponomy (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.

As GermaNet encodes all types of relation that are of interest for our project (synonymy, antonymy, hypernymy, and co-hyponymy)[2], we decided to choose it as primary source for our data sets. However, it is important to be aware of the fact that GermaNet is largely based on manually compiled sources such as thesauri, which tend to list the most salient semantic relations between words. This means that the inclusion of an entry often depends on the subjective decision of the lexicographer. Nevertheless, GermaNet is still the largest database of its kind for German, and we therefore decided to use it as starting point for our dataset.

4. Dataset I: SemrelTargets

4.1. Design

The purpose of collecting Dataset I was to acquire a broad range of lexical items which could be used as targets in generating semantically related word pairs on the one hand (cf. Section 6), and as targets for the automatic extraction of SemrelNets on the other, to create a coherent set of semantic relation pairs (to be described in Section 5). The targets were sampled randomly from GermaNet following four selection criteria. Three of these criteria were based on information available in GermaNet (part of speech, semantic category, and number of senses). A fourth criterion, corpus frequency, was established externally, since (unlike in WordNet) frequency information is not available in GermaNet. Also, we preferred to rely on larger corpus resources for frequency estimation. With no sense-tagged corpus available for German, we acquired type frequencies from a large lemmatised corpus of German (sdeWaC-3[3]). This means that lexical units (corresponding to word senses) were sampled from GermaNet according to the frequency of the corresponding lemma, and not according to the frequency of the sense itself. For polysemous targets, the frequency provided therefore subsumes the target's associated senses and semantic categories.

We used a stratified sampling procedure where for each of the three parts of speech *adjective*, *noun*, and *verb*, 99 targets were sampled randomly (but proportionally) from the following groups:

1. **Semantic categories:** *16* for adjectives, *23* for both nouns and verbs
2. **Three polysemy classes:** I) *Monosemous*, II) *Two* senses, and III) *More than two* senses
3. **Three frequency classes[4]:** I) *Low* (200–2,999), II) *Mid* (3,000–9,999), and III) *High* ($\geq$10,000)

Initially, for each part of speech, the number of lexical units required from each semantic category was established (proportionally to the total number of lexical units in the respective category), which in turn were distributed proportionally across the three polysemy classes and the three frequency classes[5]. Lexical units matching these criteria were then drawn randomly from GermaNet to populate the data set.

4.2. Results and discussion

Table 1 illustrates the choice of adjectives from the semantic categories "Gefühl" ('feeling') and "Verhalten" ('behaviour'). The former contains 7.38% of all adjectives in GermaNet (633 out of 8582). Correspondingly, a total of 7 adjectives (7.38% of 99) was drawn from this category to be included in the SemrelTargets dataset, and distributed proportionally across the nine sense and frequency classes. Similarly, the category "Verhalten" contains 13.76% of all adjectives (1181 out of 8582), from which 14 were sampled for our dataset, shown in the column labelled 'Verhalten'.

[2]GermaNet 6.0 contains a total of 74,945 hypernymy relations, and 1,587 antonymy relations.

[3]http://wacky.sslmit.unibo.it/doku.php?id=corpora

[4]Type frequency in sdeWaC-3 (Total size: 0.88 billion words)

[5]The class membership thresholds for polysemy and frequency were set manually.

Table 2 shows the distribution of polysemy in the dataset. Since polysemy classes I and II are defined to contain lexical units with exactly 1 and 2 senses, respectively, one third (33) of all selected targets for each word class are monosemous, and another third (33) have two senses. Table 2 shows the number of senses of the remaining 33 lexical units randomly sampled for each word class. The results indicate that the number of lexical units in GermaNet rapidly decreases for sense inventories greater than 3.

Senses	Adj	N	V
1	33	33	33
2	33	33	33
3	29	24	14
4	1	6	7
5	0	1	5
6	1	1	3
7	1	1	2
8	0	0	0
9	0	0	2
10	1	0	0

Table 2: Number of senses selected for each word class

Finally, Figure 2 shows that the sampled data conforms to commonly assumed models of sense frequency distributions: The more senses a lexical unit has, the larger its type frequency in corpus data. Thus, the average frequency of lexical units with one sense is 10,255, while the frequency values of lexical units with two senses and three or more senses are 18,257 and 37,479, respectively.

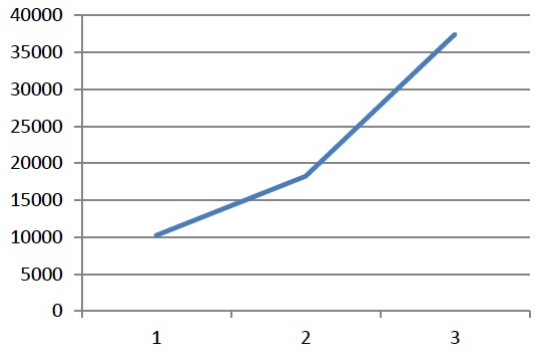

Figure 2: Average frequency per sense class (1=Monosemous, 2=Two senses, 3=More than two senses)

5. Dataset II: SemrelNets

5.1. Design

The targets generated in the previous section were used to build a second dataset containing semantically related word pairs drawn from GermaNet. The goal was to include examples of the following four major types of paradigmatic semantic relations:

1. Antonymy (ANT)
2. Synonymy (SYN)
3. Hypernymy (HYP)
4. Co-Hyponymy (COHYP)

Instead of drawing random relations from GermaNet for each of the input targets, a more sophisticated approach was taken: For each input target, a semantically coherent 'mini-network' of semantic relations was constructed using the target lexical unit (referred to as *t*) as starting point. These interconnected 'SemrelNets' aim to capture a sample of the semantic neighbourhood of *t* (in terms of synonymy, hypernymy, and co-hyponymy), as well as of its opposing one, that is, the neighbourhood of a concept that is opposite in meaning to *t*. In practice, this means that a SemrelNet N typically has the following characteristics:

- N contains a maximum of eight relations (two instances of each type): {ANT1, ANT2, SYN1, SYN2, HYP1, HYP2, COHYP1, COHYP2}.
- N contains two subnets $\{N_1, N_2\}$, where N_1 clusters around the node containing the target lexical unit *t*, while N_2 clusters around a lexical unit which stands in an antonym relation (ANT1) to *t*.
- N_1 typically contains {SYN1, HYP1, COHYP1}, while N_2 contains {SYN2, HYP2, COHYP2}.

A schematic representation of a SemrelNet N is shown in Figure 3. In this example, the boxes labelled *t* and *a1* represent the core nodes of N, and are related via antonymy (ANT1). A second antonymy link (ANT2) is chosen such that it links the synonym of *t* (i.e., *s1* as SYN1) and the synonym of *a1* (i.e., *s2* as SYN2). The antonym-synonym configuration is completed by a hypernym and a co-hyponym of the core nodes *t* and *a1*. Figure 4 shows an actual example from our data illustrating the type of SemrelNet schematised in Figure 3.

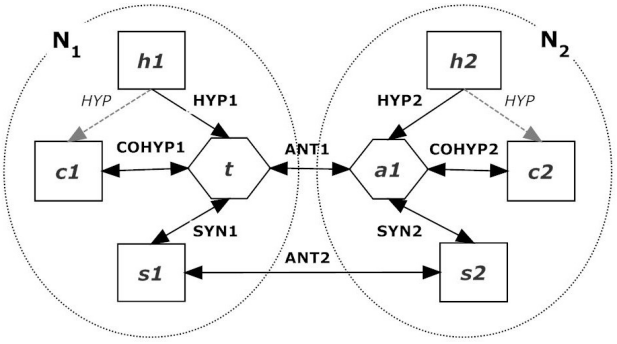

Figure 3: Schematic representation of a SemrelNet

We designed an algorithm for building SemrelNets from target lexical units in GermaNet, of which we provide an overview in the following paragraphs. One important consideration in designing the nets was to find an appropriate balance between network density and random choice of members. For our purposes, SemRelNets with higher density (i.e. with a small number of highly connected nodes) are preferable to more open networks with a larger number of nodes, as the former allows a more principled investigation of the semantic relations of specific lexical items (in particular, the input target), and their perception. For

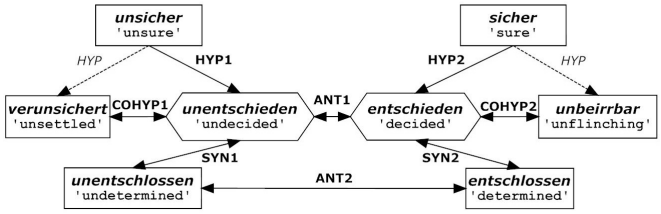

Figure 4: SemrelNet for target *unentschieden* ('undecided')

example, we assume that some paradigmatic semantic relations are more easily distinguished or confused than others, e.g., synonymy is assumed to be easily confused with hypernymy, while antonymy is assumed to be easily confused with co-hyponymy. This is to be confirmed by our experiments. On the other hand, the choice of related nodes should be as random as possible to avoid a bias towards selecting highly connected nodes in GermaNet (which may represent lexical units of higher frequency and/or higher prominence in the lexicographer's mental lexicon). The algorithm tries to take this into account by employing four main methods of locating suitable relation nodes in GermaNet, ordered here according to priority:

- Method 1: Direct-motivated
- Method 2: Direct-random
- Method 3: Indirect-random
- Method 4: Broken-random

In the schematic example shown in Figure 3, the relations ANT1 and ANT2, as well as SYN1 and SYN2, are selected via the ***direct-motivated*** method (Method 1). The goal of this method is to locate a direct antonym *a1* of *t*, which has a synonym *s1* (or hypernym/hyponym *h1*) which is itself in an antonymy relation with a synonym *s2* (or hypernym/hyponym *h2*) of *a1*. The other relations in the network are then chosen via the ***direct-random*** method (Method 2), where the algorithm tries to find nodes in the GermaNet network that are directly attached to *t* and *a1* via the required relation types (in this case HYP1, COHYP1 and HYP2, COHYP2). If several nodes are available, a random choice is carried out. Thus, in Figure 4, the synonymy relations SYN1 and SYN2 as well as the antonymy relations ANT1 and ANT2 were established via the *direct-motivated* method in our algorithm, while HYP1, COHYP1 and HYP2, COHYP2 were established randomly via the *direct-random* method.

Methods 1 and 2 aim to maximise network density: By choosing synonyms of *t* and *a1* that are themselves related via antonymy, Method 1 aims to increase the density of the resulting net. On the other hand, Method 2 also works towards a close-knit net by choosing relations that are directly attached to *t* and *a1*. In addition, a special procedure applies to the direct-random choice of hypernyms and co-hyponyms: To increase the connectivity of the SemrelNet, preference is given to co-hyponyms and hypernyms of the target (cf. *c1* and *h1* in Figure 3) which are themselves related via a hypernymy relation (as is the case in Figure 4, where the dotted lines indicate a hypernymy relation). For this purpose, the algorithm first chooses a random co-hyponym of the target (but excluding lexical units which are simultaneously synonyms or antonyms of the target), and then includes the corresponding hypernym (if several are available, a random one is selected). Reversing the procedure by randomly choosing a hypernym first and then selecting one of its hyponyms as co-hyponym of the target would result in low probabilities for co-hyponyms with many siblings. Finally, while artificial concepts (cf. Section 3) are generally excluded from consideration as SemrelNet members, they are allowed as common hypernym of a target and its co-hypernym. Therefore, in cases where the corresponding hypernym turns out to be an artificial node in GermaNet, the co-hyponym is still selected, but another (non-artificial) hypernym or hyponym is randomly determined for the HYP relation. Figure 5 shows an example of a SemrelNet where COHYP2 involves an artificial common hypernym (*geschwindigkeitsspezifisch*, 'speed-specific'). HYP2 was determined in a second step via the *direct-random* method, which located a direct hyponym of *langsam* ('slow'): *schleppend* ('sluggish').

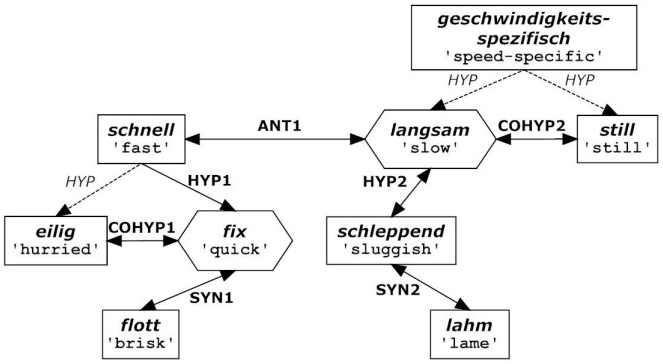

Figure 5: SemrelNet for target *fix* ('quick')

Figure 5 illustrates the situation where the first two methods fail because *t* does not have a direct antonym *a1*. This is where Method 3 (***indirect-random***) comes in: If no direct relations are available, the algorithm checks if any of the already existing nodes in the respective subnetwork (i.e. nodes which have already been filled by previous methods) are involved in one or more relations of the required type. If a match is found, a randomly-chosen relation and its associated node are added to the SemrelNet. The order in which existing nodes are checked is synonyms (1.), hypernyms/hyponyms (2.), and finally co-hyponyms (3.). For example, while SYN1, HYP1, and COHYP1 in Figure 5 were chosen via the direct-random mode, both ANT1 (attaching to the hypernym of the target of N_1, *schnell* 'fast') and SYN2 (attaching to the hyponym of the target of N_2, *schleppend* 'sluggish') were retrieved via the indirect-random method.

Finally, a back-off strategy was implemented to check for relations involving nodes that are directly connected to the target but not included as existing nodes in the given SemrelNet (Method 4, ***broken-random***). This means that there is no existing path in the network between the target and the (randomly selected) node, as illustrated in Figure 6. Here, SYN2 was chosen via the broken-random mode: The lexi-

cal unit *versauen* (‘to blow sth.’) has been marked as synonym of *vermasseln* (‘to mess up’), which is a hyponym of the target *durchfallen* (‘to fail (a test/exam)’). However, this hypernymy relation is not itself part of the network N, resulting in a broken path between *durchfallen* and *vermasseln*.

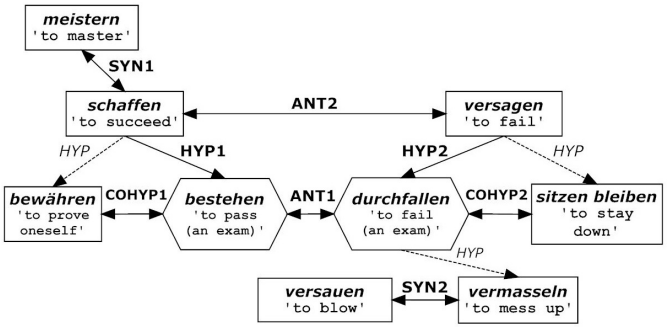

Figure 6: SemrelNet for target *bestehen* (‘to pass’)

Depending on the density of the network surrounding *t*, any number of nodes and associated relations in N may be blank: For example, if *t* and *a1* had no synonyms, the nodes *s1* and *s2*, as well as the relations SYN1/SYN2/ANT2, would be missing from the diagram in Figure 3. Similarly, if no antonym *a1* can be found for the members of N_1, subnet N_2 remains completely blank.

5.2. SemrelNet extraction tool

The algorithm described in the previous section has been implemented in Java and directly draws on the latest version of the GermaNet Java API (6.0.1)[6], which provides access to all information in GermaNet 6.0. A number of new classes and methods were implemented centering around the new concept ‘SemrelNet’. Instances of the SemrelNet class consist of a number of nodes (representing any participating lexical units in the SemrelNet, such as *s1*, *s2*, *h1*, *h2*, etc. as shown in Figure 3) and relations (linking a pair of nodes). For example, in the SemrelNet for target *unentschieden* (‘undecided’, cf. Figure 4), node *t* is realised by the lexical unit *unentschieden*, *s1* is realised by *unentschlossen* (‘undetermined’), and SYN1 links *t* and *s1*. In addition to their function in the net and the lexical unit which realises them, instances of the node class further record information about their position in the SemrelNet, relative to the target node *t*. For instance, node *s1* is typically involved in a synonymy relation within N_1, but due to the indirect-random and broken-random methods (cf. Section 5.1) it may appear in various positions within the subnet. For example, in Figure 6, *s1* (realised by *meistern*, ‘to master’) is an indirect synonym of *t*, being attached to the hypernym *h1* of *t*.

Table 3 provides an overview of the naming conventions used for node positions in a given SemrelNet, while Figure 7 shows the SemrelNet for target *bestehen* (‘to pass’) (cf. Figure 6) with added node labels of the format ‘function: position’. The labels show, for instance, that the node containing the lexical unit *sitzen bleiben* (‘to stay down (at school)’) has the function ‘c2’ (co-hypernym 2) and the position ‘cat’ (‘co-hypernym of antonym of t’). The position information on the *s2* (synonym 2) node with lexical unit *versauen* (‘to blow sth.’) indicates that there is a ‘broken’ path between *a1* and its hyponym *vermasseln* (‘to mess up’): In this case, ‘sUat’ reads as ‘synonym of broken hyponym of antonym of t’). Providing position information as shown in Table 3 is crucial for the graphical visualisation of SemrelNets.

Position	Read as...
t	target
sx / Sx	synonym / ‘broken’ synonym of x
ax / Ax	antonym / ‘broken’ antonym of x
ox / Ox	hypernym / ‘broken’ hypernym of x
ux / Ux	hyponym / ‘broken’ hyponym of x
cx / Cx	co-hyponym / ‘broken’ co-hyponym of x

Table 3: Naming conventions for SemrelNet node positions

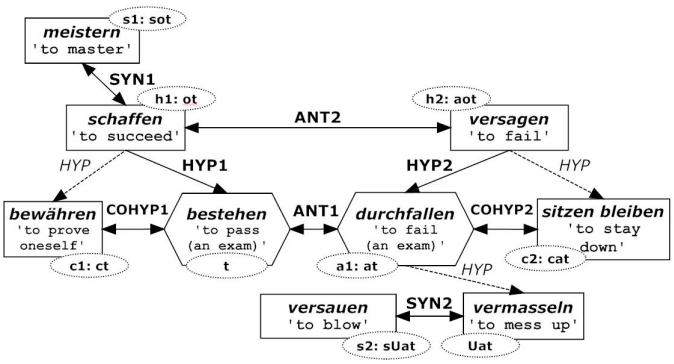

Figure 7: SemrelNet for target *bestehen* (‘to pass’) with added node position labels

The SemrelNet extraction tool produces two kinds of output: a simple text-based format (Figure 8) and XML format (Figure 9). In addition to listing the nodes and relations included in the nets, the output also provides information in terms of the GermaNet-IDs of all lexical units (attribute ‘id’ in Figure 9), and for each SemrelNet information about the target’s part of speech (attribute ‘pos’), semantic category (‘cat’), number of senses (‘senses’), corpus frequency (‘freq’), depth in the GermaNet hierarchy (‘depth’), and an overview of the completeness of the net (‘statsCode’ and ‘completeness’).

The SemrelNet extraction tool is freely available[7] and can be run on the whole of GermaNet, or on a selected list of lexical units. Due to the random methods included in the algorithm the resulting SemrelNets may be different when the tool is re-run several times on the same input data.

5.3. Results and discussion

This subsection intends to give an overview of the results of running the tool on the SemrelTargets dataset described

[6] http://www.sfs.uni-tuebingen.de/lsd/javadoc6.0/index.html

[7] http://www.ims.uni-stuttgart.de/projekte/semrel/resources.html

```
bestehen_76346 | V | Gesellschaft | 7 | 334228 | [4] | 1-1-1-1-1-1-1-1 | 8
ANT1: Dir-Motiv-AntR1 t:at | bestehen_76346 : durchfallen_76372
SYN1: Indir-Rand-Syn-T ot~sot | schaffen_76343 ~ meistern_76344
HYP1: Dir-Motiv-HypR1 t<ot | bestehen_76346 < schaffen_76343
COHYP1: Dir-Motiv-HypR1 t--ct | bestehen_76346 -- bewähren_76365
    ----> CH: schaffen_76343
----
ANT2: Dir-Motiv-HypR1 ot:oat | schaffen_76343 : versagen_76369
SYN2: Broken-Rand-Syn-A Uat~sUat | vermasseln_76375 ~ versauen_76376
HYP2: Dir-Motiv-HypR1 at<oat | durchfallen_76372 < versagen_76369
COHYP2: Dir-Motiv-HypR1 at--cat | durchfallen_76372 -- sitzen bleiben_76373
    ----> CH: versagen_76369
```

Figure 8: Text-based output of SemrelNet extraction tool

```
<relnet t="bestehen" id="76346" pos="V" cat="Gesellschaft" senses="7"
freq="334228" depth="4" statsCode="1-1-1-1-1-1-1-1" completeness="8">
<relation type="ANT1" rule="Dir-Motiv-AntR1">
<lu pos="t" id="76346">bestehen</lu>
<lu pos="at" id="76372">durchfallen</lu></relation>
<relation type="SYN1" rule="Indir-Rand-Syn-T">
<lu pos="ot" id="76343">schaffen</lu>
<lu pos="sot" id="76344">meistern</lu></relation>
<!-- remaining relations omitted for space reasons -->
</relnet>
```

Figure 9: XML output of SemrelNet extraction tool

in Section 4. Table 4 shows the size of the SemrelNets generated for the individual word classes. Complete nets (i.e. nets containing two instances of each of the four semantic relations ANT, SYN, HYP, and COHYP) are achieved for two thirds of all input adjectives (66), one third of verbs (32), but only for around one fifth of all input nouns (18). This is due to the fact that fewer nouns are involved in antonymy relations, which results in a large number of missing subnets N_2 (cf. Figure 3). As a consequence, the noun dataset contains a large number of SemrelNets of size 3 (59 altogether), typically containing the relations SYN1, HYP1, and COHYP1 (Figure 10).

Relations per net	Adj	N	V	All
1	0	0	0	0
2	4	4	4	12
3	21	59	47	127
4	0	0	0	0
5	1	0	0	1
6	1	0	1	2
7	6	18	15	39
8	66	18	32	116

Table 4: Number of relations per SemrelNet

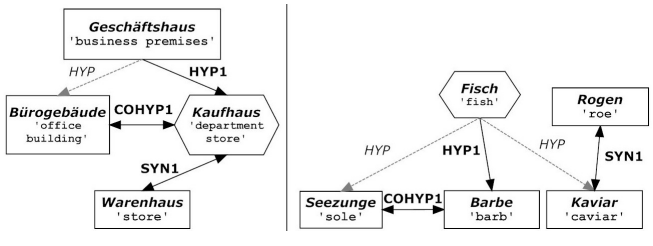

Figure 10: Examples of SemrelNets with three relations

With the exception of one example, which contains SYN1 and COHYP1 only, all 12 SemrelNets with only two relations include a HYP1 and COHYP1 relation (examples are shown in Figure 11). This is due to GermaNet's focus on the hypernymy hierarchy, which means that, generally, hypernyms and co-hyponyms are available for most lexical entries. All SemrelNets with three relations are of the type SYN1-HYP1-COHYP1 (as illustrated in Figure 10).

There are no SemrelNets with 4 relations, which again follows from GermaNet's structure as hypernym hierarchy: As soon as an antonym relation ANT1 is available, the paired lexical unit (referred to as *a1* in Figure 3) is likely to be involved in a hypernymy (HYP2) and/or co-hypernymy (COHYP2) relation. In other words, if a SemrelNet contains four relations, it will automatically contain a minimum of five relations altogether. Finally, it is worth noting that most instances of SemrelNets with seven relations (36 of 39) are missing an antonym relation, because antonymy is underrepresented across word classes (Figure 12).

Table 5 lists the total number of relation types included in the dataset. As expected, with the exception of one adjective, all input targets have SemrelNets which contain HYP1 and COHYP1 relations. The table further shows that an 'opposing' subnet N_2 exists for 74 adjectives (75%), 36 nouns (36.4%), and 48 verbs (48.5%, cf. row 'ANT1'). All

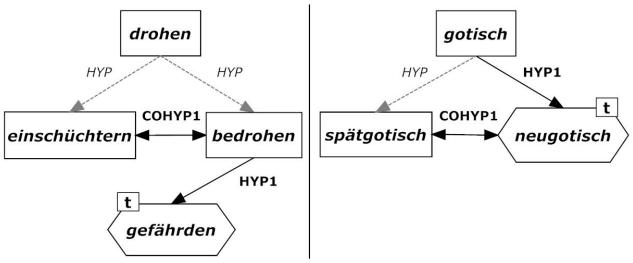

Figure 11: Examples of SemrelNets with two relations

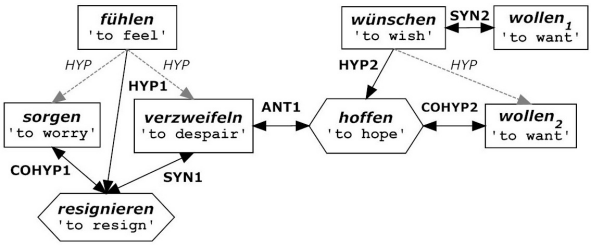

Figure 12: Example of SemrelNet with seven relations

N_2 include HYP2 and COHYP2 relations (with the exception of one adjective). Almost equally complete are the synonym relations: Only four adjectives, four nouns, and six verbs have no SYN1 relation in their network, and nearly all SemrelNets with a subnet N_2 also include a SYN2 relation (73 of 74 for adjectives, 36 of 36 nouns, and 47 of 48 verbs). The relation that fares worst in these statistics is ANT2: Only 18.2% (18) of noun targets, and 34.3% (34) of verb targets have a SemrelNet which includes ANT2. As noted above, this is due to the fact that (particularly for nouns) only a small number of antonym relations are encoded in GermaNet, and the chances of finding two of them within the same SemrelNet are therefore low. The situation is slightly better for adjective targets: Here, 67.7% (67) of SemrelNets contain an ANT2 relation. This is not surprising, since antonymy is considered the central organising principle for the adjectives in WordNets (Miller, 1990).

Relation	Adj	N	V	Total
ANT1	74	36	48	158
SYN1	95	95	93	283
HYP1	98	99	99	296
COHYP1	99	99	99	297
ANT2	67	18	34	119
SYN2	73	36	47	156
HYP2	73	36	48	157
COHYP2	73	36	48	157
TOTAL	652	455	516	1623

Table 5: Total number of relation types per word class

Finally, Table 6 gives an overview of how often the four extraction methods (described in Section 5.1) were employed in running the SemrelNet extraction tool on the input. The numbers show that the *direct-random* method is the most frequent by far, generating 66.3% of all relations (1076 of 1623). This supports the overall goal of making SemrelNets as random as possible, while still maintaining close density within the nets (by attaching relations directly to the target nodes). In contrast, the *direct-motivated* rules, whose aim is to maximise connectivity by detecting a second antonymy link between subnets N_1 and N_2, were only triggered 110 times for all word classes, being most frequently used for adjectives (71 times). The second most frequent method in all word classes is the *indirect-random* one with 15.5% for adjectives (101/652), 14.1% for nouns (64/455), and 16.1% for verbs (83/516). The use of this method results in a lower density of the net, as the selected relations are only indirectly attached to the target. However, the method still supports connectivity of the nets, as the relations are attached to other existing nodes in the net. The back-off strategy, in which so-called *broken-random* relations are considered, is used least frequently among the random relations for all word classes, with 10.0% of all adjective relations (65), 11.0% of all noun (50), and 11.4% of all verb relations (59) having been triggered by this method. Of the resulting broken-random relations included in the dataset, more than half are antonyms (56.9%, 99/174), 36.8% synonyms (64/174), and 6.3% hypernyms (11/174).

Method	Adj	N	V	Total
Direct-motivated	71	8	31	110
Direct-random	409	332	335	1076
Indirect-random	101	64	83	248
Broken-random	65	50	59	174
Other	6	1	8	15
TOTAL	652	455	516	1623

Table 6: Number of methods employed per word class

6. Current and future work

The datasets described in the previous sections are currently being used in two types of human judgement experiments: One focusing on the generation of semantically related word pairs, and the other on human rating of the strength of semantic relations. Both experiments are hosted on Amazon Mechanical Turk (MTurk)[8].

The purpose of the first experiment is to gather human associations for each type of semantic relation. That is, for each lexical unit in the SemrelTargets dataset, participants are asked to generate one synonym, one antonym, one hypernym, and one co-hyponym. In order to avoid confusion between the different types of relation, the data is presented to participants in bundles of 11 words (or 11 "HITs", as individual decision tasks are called in MTurk) to be assessed for the same type of relation (e.g. finding antonyms for each of the 11 words). The goal is to receive associations from at least 10 different participants for each target. To make sure that the data is dealt with properly, and to exclude non-native speakers of German, each set of 11 HITs includes two examples of non-words, which should be recognised

[8] www.mturk.com

as such by native speakers of German (e.g. *Blapselheit*, *gekortiert*). If not, the whole set is excluded.
In the second experiment, participants are presented with word pairs included in the SemrelNets dataset, and asked to rate their degree of synonymy, antonymy, etc. on a scale between 0 and 5, plus an option for marking unknown words. Again, to avoid confusion between the different types of relation, each bundle of 14 HITs is rated according to one specific relation at a time. Each bundle contains:

1. 3 focus-relation pairs (i.e. the relation under consideration)
2. 9 other-relation pairs (i.e. 3 pairs each from the other three relations)
3. 2 test pairs (involving one nonsense-word)

Once the experiments are completed, each word pair in the SemrelNets database will have received 10 ratings each for their degree of synonymy, antonymy, hypernymy, and co-hyponymy.

7. Conclusion

This paper described the design and compilation of a new dataset containing semantically coherent relation pairs drawn from GermaNet. The dataset consists of two parts:

1. Three sets of 99 lexical units drawn from the three major word classes adjectives, nouns, and verbs, using a stratified sampling technique to balance the datatset for semantic category, polysemy, and type frequency ('SemrelTargets'); and
2. Three sets of 99 semantically coherent subnets clustering around the SemrelTargets, and consisting of a total of 1623 paradigmatic semantic relation pairs ('SemrelNets').

The data is currently being used in two human judgement experiments, in which (1) new relation pairs are generated from the set of SemrelTargets, and (2) word pairs in the SemrelNets are rated for the strength of the semantic relations holding between them. The dataset thus promises to be an important resource not only for research in computational linguistics, but also for studies in theoretical linguistics and psycholinguistics.

8. References

Sharon A. Caraballo. 2001. *Automatic Acquisition of a Hypernym-labeled Noun Hierarchy from Text*. Ph.D. thesis, Brown University.

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

James Curran. 2003. *From Distributional to Semantic Similarity*. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding Contradictions in Text. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1039–1047, Columbus, OH.

Philip Edmonds and Graeme Hirst. 2002. Near-Synonymy and Lexical Choice. *Computational Linguistics*, 28(2):105–144.

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.

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.

Sanda M. Harabagiu, Andrew Hickl, and Finley Lacatusu. 2006. Negation, Contrast and Contradiction in Text Processing. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 755–762, Boston, MA.

Marti Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France.

Jay J. Jiang and David W. Conrath. 1997. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, pages 19–33.

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.

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2001. Identifying Synonyms among Distributionally Similar Words. In *Proceedings of the International Conferences on Artificial Intelligence*, pages 1492–1493, Acapulco, Mexico.

Cupertino Lucerto, David Pinto, and Héctor Jiménez-Salazar. 2004. An Automatic Method to Identify Antonymy Relations. In *Proceedings of the IBERAMIA Workshop on Lexical Resources and the Web for Word Sense Disambiguation*, pages 105–111, Puebla, Mexico.

John Lyons. 1977. *Semantics, Volume II*. Cambridge University Press, Cambridge, UK.

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

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.

George A. Miller, editor. 1990. *WordNet: An On-line Lex-*

ical Database, volume 3 (4). Oxford University Press. Special Issue of the International Journal of Lexicography.

Jane Morris and Graeme Hirst. 2004. Non-Classical Lexical Semantic Relations. In *Proceedings of the HLT Workshop on Computational Lexical Semantics*, Boston, MA.

M. Lynne Murphy. 2003. *Semantic Relations and the Lexicon*. Cambridge University Press.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, Sydney, Australia.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, pages 448–453, San Francisco, CA.

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

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning Syntactic Patterns for Automatic Hypernym Discovery. *Advances in Neural Information Processing Systems*, 17.

Peter D. Turney. 2006. Expressing Implicit Semantic Relations without Supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 313–320, Sydney, Australia.

Torsten Zesch and Iryna Gurevych. 2006. Automatically Creating Datasets for Measures of Semantic Relatedness. In *COLING/ACL 2006 Workshop on Linguistic Distances*, pages 16–24, Sydney, Australia.