# Enriching GermaNet with verb-noun relations – a case study of lexical acquisition

**Lothar Lemnitzer, Holger Wunsch and Piklu Gupta**

Seminar für Sprachwissenschaft,
Universität Tübingen,
Germany,
{lothar,wunsch,gupta}@sfs.uni-tuebingen.de

**Abstract**

In this paper we will focus on the lexical-semantic relations in the German wordnet GermaNet. It has been shown that wordnets suffer from the relatively small number of relations between their lexical objects. It is assumed that applications in NLP and IR, in particular those relying on word sense disambiguation, can be boosted by a higher relational density of the lexical resource. We report on research and experiments in the lexical acquisition of a new type of relation from a large annotated German newspaper corpus, i.e. the relation between the verbal head of a predicate and the nominal head of its argument. We investigate how the insertion of instances of this relation into the German wordnet GermaNet affects the overall structure of the wordnet as well as the neighbourhood of the nodes which are connected by an instance of the new relation.

## 1. Introduction

Wordnets are a valuable lexical-semantic resource used with many NLP applications and techniques (cf. Morato et al. (2003)). The main characteristics of wordnets are the organisation of lexical units into synsets and the connection of both lexical units and synsets by lexical-semantic relations. In this paper we will focus on the lexical-semantic relation types. The types of lexical semantic relations which can be found in nearly all wordnets are a) synonymy; b) antonymy; c) hypernymy / hyponymy; d) holonymy / meronymy; e) troponymy ; f) causation; g) entailment; h) pertainymy. We call these relations the classical relation types. Many NLP applications which use wordnets as a lexical resource draw on information about the semantic relatedness or similarity of lexical units which co-occur in documents. These applications view wordnets as graphs in which the synsets are the nodes and the relations between these objects are the edges. Semantic relatedness or similarity between synsets and lexical units (in the following: lexical objects) is measured by the length of the shortest path which connects the lexical objects[1]. It has been shown recently that wordnets suffer from the relatively small number of relation instances between their lexical objects (cf. Boyd-Graber et al. (2006)). It is assumed that applications in NLP and IR, in particular those relying on word sense disambiguation, can be boosted by a lexical-semantic resource with a higher relational density and, consequently, shorter average paths between the lexical objects. In this paper, we report on research in the lexical acquisition of a new type of relation from a large annotated German newspaper corpus. We focus on the relation between the verbal heads of predicates and the nominal heads of their arguments. We investigate how the insertion of instances of this relation into the German wordnet GermaNet affects a) the overall structure of the wordnet and b) the neighbourhood of the nodes which are connected by an instance of the new relation. In particular, we will observe the decrease in the sum total of all path lengths connecting the nodes. To achieve this, we calculate the shortest paths between any two synsets and present the sum total as well as the distribution of these path lengths. We compare the measures for the original wordnet and the wordnet with new relation instances added. The impact of the new relation instances on the sum and distribution of path lengths serves as a benchmark for the efficiency of several acquisition methods.

We expect the introduction of new, non-classical relations between concepts to have a positive impact on applications which draw on measurements of semantic relatedness between concepts.

The rest of the paper is organised as follows: we start with an overview of related work; in section 3. we describe the corpus which we have used for our acquisition experiments, in section 4. the acquisition methods are described. Section 5. is devoted to the experiments with the extended GermaNet and their outcomes. We finish the paper with a section in which we draw conclusions and outline future work.

[1] Budanitsky and Hirst (2006) give an overview of methods by which the shortest path between any two lexical objects has been calculated.

## 2. Related Work

The work on which we report in our paper is an example of acquisition of lexical-semantic descriptions with the aim of structurally enriching a lexical-semantic resource. Research in the acquisition and integration of new synsets aims to reduce the amount of time-consuming and error-prone manual work required to extend these resources. Snow et al. (2006) and Sang (2007) present highly efficient approaches to this task. They exploit the fact that taxonomic relations between lexical objects are reflected by distributional patterns of these lexical objects in corpora. This kind of research, however, is not in the scope of this paper. Instead, we deal with the introduction of relation instances between synsets which are already included in the wordnet, and in particular with instances of a new type of relation. This relation type connects verbal predicates and their nominal arguments.

Research in the (semi-)automatic detection and integration of relations between synsets has boomed in recent years. These activities can be seen as a response to what Boyd-Graber et al. (2006) identify as a weakness of the Princeton WordNet: "WordNet, a ubiquitous tool for natural language processing, suffers from sparsity of connections between its component concepts (synsets)." Indeed, the number of 68,000 relation instances connecting 53,312 synsets and 76,563 lexical units in GermaNet is surprisingly low and needs to be increased. We assume that similar ratios between objects and relations characterise many wordnets.

An entire task of the latest SEMEVAL competition has been dedicated to the detection and classification of semantic relations between nominals in a sentence (cf. Girju et al. (2007)). This line of research, however, is targeted at the detection of classical lexical-semantic relations like hyperonymy and at relations between words of the same part of speech (i.e. nouns). We intend to introduce a new syntagmatic relation which relates verbal predicates with their nominal arguments.

Some effort has been made to introduce non-classical, cross-category relations into wordnets. Boyd-Graber et al. (2006) introduce a type of relation which they call "evocation". This relation expresses that the source concept as a stimulus evokes the target concept. In other words, this is a mental relation which cuts across all parts of speech. This makes the approach different from ours, since we use corpus data instead of experimental data and we acquire what is in the texts rather than what is in the human mind. The relation we introduce is syntactically motivated, which is not the case in the experiment on which Boyd-Graber et al. report.

Amaro et al. (2006) intend to enrich wordnets with abstract predicate-argument structures, where the arguments are not real lexical units or synsets but rather abstract labels like INSTRUMENT. They aim at a lexical-semantic resource which supports the semantic component of a deep parser. Therefore they introduce a level of abstraction in the categorisation of the arguments. This is not what we intend to do.

Yamamoto and Isahara (2007) extract non-taxonomic, in particular thematic relations between predicates and their arguments. They extract these related pairs from corpora by using syntactic relations as clues. In this respect their work is comparable to ours. Also their aim, i.e. improving the performance of information retrieval systems with this kind of relation, is comparable to ours. However, they do not use the extracted word sets to include them in a wordnet.

Closest to ours is the work of Bentivogli and Pianta (2003). Their research is embedded in the context of machine translation. Seen from this perspective, the almost exclusive representation of single lexical units and their semantic properties is not satisfying. They therefore propose to model the combinatoric idiosyncrasies of lexical units by two new means: a) the phraset as a type of synset which contains multi-word lexical units and b) syntagmatic relations between verbs and their arguments as an extension of the traditional paradigmatic relations. Their work, however, focuses on the identification and integration of phrasets. They only resort to syntagmatic relations where the introduction of a phraset would not otherwise be justified. We take the opposite approach in that we focus on the introduction of instances of the verb-argument relation and resort to the introduction of phrases only in those cases where it is not possible to ascribe an independent meaning to one of the lexical units (cf. section 4.).

## 3. The corpora

For the acquisition experiments we use the *Tübingen Partially Parsed Corpus of Written German* (TüPP-D/Z). This corpus contains approximately 11.5 million sentences and 204,661,513 lexical tokens. It has been automatically annotated using the cascaded finite state parser *KaRoPars* (cf. Müller (2004)). Four levels of syntactic constituency are annotated: a) the lexical level, b) the chunk level, c) the level of topological fields, and d) the clausal level. Parse trees are quite flat in TüPP-D/Z. Due to limitations of the finite state parsing model, the attachment of chunks remains underspecified. Major constituents are annotated with grammatical functions (cf. fig. 1. The example sentence translates to: *We need to sell the villas in order to pay the young scientists.*). The relevant information for the extraction of verb-object pairs, most importantly the annotation of topological fields and of noun chunks with grammatical functions, is present with sufficient accuracy. From the example above, the pairs *brauchen, Villenverkauf* and *bezahlen, Nachwuchs* will be extracted.

The results of the automatic linguistic analysis, however, have not been corrected manually, due to the size of the corpus. Therefore we have to choose an acquisition method which is not sensitive to errors in the annotation.

## 4. Acquisition methods used

Starting from the linguistically analysed and annotated corpus which we have described above, we extracted two types of syntactically related word pairs: a) verb-subject (e.g. *untersuchen, Arzt – examine, doctor*) and b) verb-direct object (e.g. *finden, Weg,– find, way*).

While the spectrum of possible subjects of a verb turned out to be much broader and heterogeneous, verb-object pairs were more readily identifiable and recurrent. Even in scenarios in which associations are arrived at on the basis of evocation, it is interesting to observe that, for instance, Schulte im Walde (2006) found a higher number of associations arrived at by humans between verbs and their direct objects than between verbs and their transitive or intransitive subjects. Therefore we focused our work on the analysis of verb-object pairs.

In order to rank the word pairs, we measured their collocational strength, which we consider to be a good indicator for their semantic relatedness. Two common measures – mutual information (MI) (Church et al., 1991) and log-likelihood ratio (often referred to as $G^2$ (Dunning, 1993)) – are used and compared in our experiments. Mutual information can be regarded as a measurement of how strongly the occurrence of one word determines the occurrence of another; it compares the probability of, for example, two words occurring together with the probability of observing them independently of one another. Log-likelihood ratio

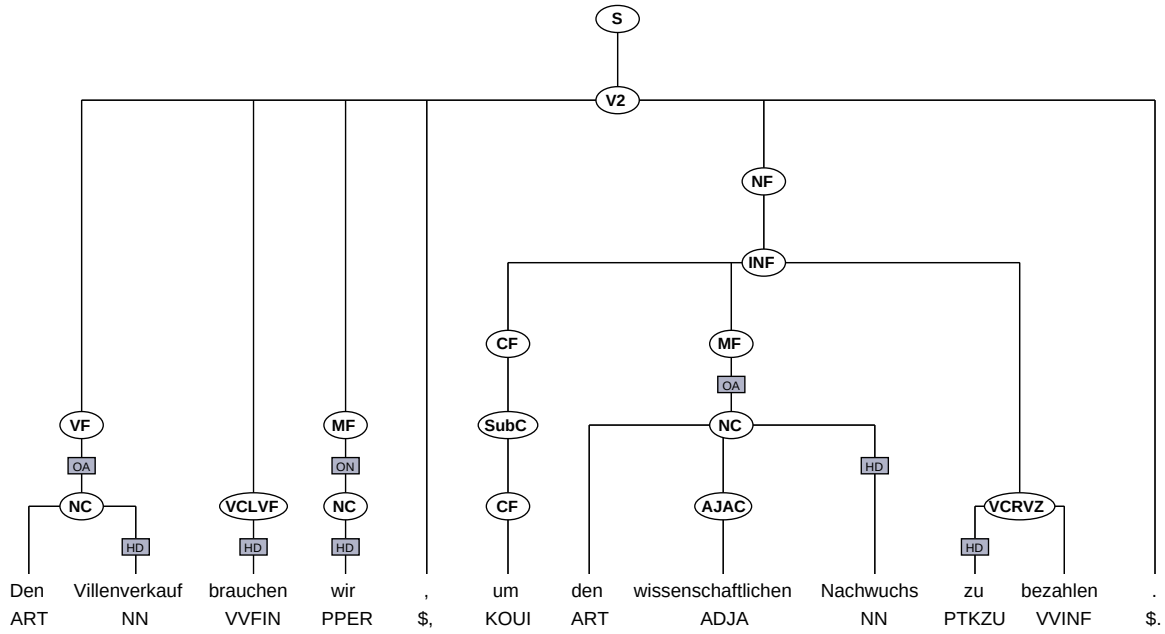


Figure 1: Parse tree of a TüPP-D/Z sentence

compares expected and observed frequencies as might be expressed in a contingency table, a 2 by 2 table where the four cell values represent frequency of word *x* occurring with word *y*, *x* and not *y*, not *x* and *y* and finally not *x* and not *y*, i.e. the number of observations where neither word appears. The benefits of log-likelihood over mutual information are outlined by Kilgarriff (1996) and by Lemnitzer and Kunze (2007) who argue that mutual information overemphasises rare terms in a corpus, whereas log-likelihood ranks frequent pairs higher, but also delivers reliable results when the occurrence of certain words or bigrams in a corpus are rare.

In order to compare both measures, we took the first 100 entries from the lists of MI-ranked and $G^2$-ranked word pairs, after having cleaned these lists, and inserted them manually into GermaNet. Before we selected these word pairs, we had to clean the lists of pairs which we did not want to insert into the wordnet. We removed a) pairs with wrongly assigned words due to errors in the linguistic annotation and b) pairs with words which did not have an entry in GermaNet. We further removed collocations and (parts of) fixed expressions as well as support verb constructions. We consider it to be inappropriate to represent (semi-)fixed expressions by relating their elements. Instead, we will encode these expressions in a way which has been sketched by Bentivogli and Pianta (2003). They encode, for example, noun-preposition-noun expressions as one unit rather than by relating their elements to each other.

The 100 highest ranked of the remaining word pairs were inserted into the wordnet manually. Inserting the word pair involved a manual disambiguation step: all words were mapped to the correct synsets. Semi-automatic insertion of the new relation instances would require reliable word sense disambiguation which is not yet available for German. In the following we report on experiments in which we calculated the global and local impact of the new relation instances.

# 5. Experiments

## 5.1. Network-global effects of adding relations

If a verb imposes strong selectional preferences on the objects it takes, the semantic fields that the verb and the object denote are usually closely related to each other as well. A good example for this are verbs of digestion, such as *eat*, which take concepts from the semantic field of *food* as their objects in the vast majority of all cases. In the first set of experiments, we examined network-global effects of adding a new relation between a verb and an object which belong to a pair that was deemed to be of good quality according to some measure, as described in the previous section. Given that the verb-object pair was good, the new relation explicitly connects two closely related semantic fields that had not been connected before, one field in the verb part, and the other field in the noun part of the network. Increasing the relational density by adding a new relation alters the global topology of the network in such a way that the total length of all shortest paths between all concepts is decreased due to the new "shortcuts" between verb and noun semantic fields. Furthermore, our hypothesis is that the better the verb-object pair, the more the total path length decreases.

The baseline for all experiments was GermaNet version 5.0. GermaNet contains multiple top nodes that correspond to the most general semantic concepts in their respective word class, called unique beginners, which are not connected to each other. We added a new artificial top node and edges that connect the unique beginners via the artificial top node. In this way, the existence of (an albeit long) path between any two verbs and nouns was guaranteed in the baseline network[2]. We conducted five experiments with five different sets of input data. Each set consisted of a list of 100 verb-object pairs (with the exception of one data set, which only

[2]By adding the artificial top node and the connecting edges, GermaNet is turned into a connected graph.

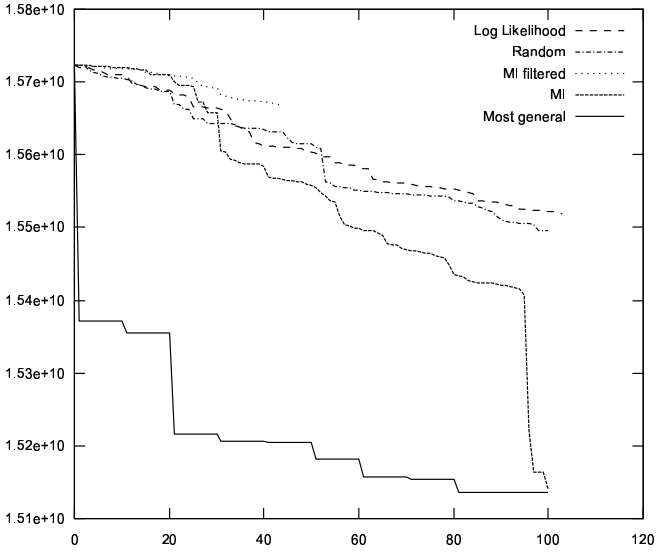


Figure 2: Plot of cumulative path lengths. The values on the X-axis are the number of pairs added, the values on the Y-axis are the cumulative path length.

contains 44 pairs). For all pairs, new relations were successively added to GermaNet. These relations connected the synset that corresponds to the verb with the synset that corresponds to the noun. After each addition, we computed the cumulative length of all shortest paths in the network.

The data set for the first experiment, **Experiment G-LogLik**, was acquired by extracting verb-object pairs from the TüPP-D/Z treebank, and then calculating the log-likelihood ratio for each pair. The 100 top ranked pairs were considered in the experiment. For **Experiment G-MI**, we used the same verb-object pairs, but this time we used the 100 pairs ranked top according to mutual information. Since this list contained many pairs that didn't make much sense intuitively, we conducted **Experiment G-MI-Filtered** with a manually filtered list of 44 plausible pairs. Additionally, we performed two baseline experiments. **Experiment G-Random** is based on 100 verb-object pairs randomly selected from GermaNet. **Experiment G-General** contains 100 verb-object pairs taken from the most general layer of synsets in GermaNet. Our hypothesis was that the cumulative path reduction would be higher when relations corresponding to semantically meaningful pairs are added to GermaNet.

Figure 2 shows that our hypothesis does not prove true. The highest reduction of cumulative path length occurs in **Experiment G-General**, while the other data sources yield largely similar results. This leads to the conclusion that on the network-global level, changes of cumulative path lengths are only determined by general effects of increased network density, but not by additional effects of the semantic relatedness of the newly connected subnets.

### 5.2. Local effects of adding relations

In this experiment, we introduced the relation instances one at a time. The settings are as follows: let $s_1$ and $s_2$ be two synsets and $R(s_1, s_2)$ the new relation instance connecting the two synsets. Further, let $SP_b$ be the shortest path between $s_1$ and $s_2$ before the insertion of $R(s_1, s_2)$ and let $SP_a$ be the shortest path between $s_1$ and $s_2$ after the insertion of $R(s_1, s_2)$. By definition, the length of the shortest path between $s_1$ and $s_2$ after the insertion of ($SP_a$) is 1 (see figure 3). We calculate the path reduction $PR_{s_1,s_2}$ as the result of $SP_b - SP_a$. We now take $S_1$ and $S_2$, the sets of all synsets which are in the two subtrees rooted by $s_1$ and $s_2$ respectively; in other words, we take all the hyponyms, the hyponyms of these hyponyms and so forth. We calculate the path reduction $PR_{s_m,s_n}$ for each pair $s_m \in S_1, s_n \in S_2$. The sum of all path reduction values is the local impact caused by the new relation instance. We calculated the sum total of the path reduction values for the 100 most highly ranked pairs according to the MI and $G^2$ statistics. Table 1 shows the average cumulative path reduction value for both statistics.

| method | average PR value |
|---|---|
| MI | 2762.04 |
| $G^2$ | 15867.38 |

Table 1: Cumulative path length reduction, average of 100 word pairs for both MI and $G^2$.

From these figures we can infer that: a) there is a considerable local impact of the new relation instances; b) the impact of the word pairs extracted by $G^2$ is much higher than that of the pairs extracted by $MI$, which was expected given that MI ranks pairs of infrequent words and therefore more specialised words higher.

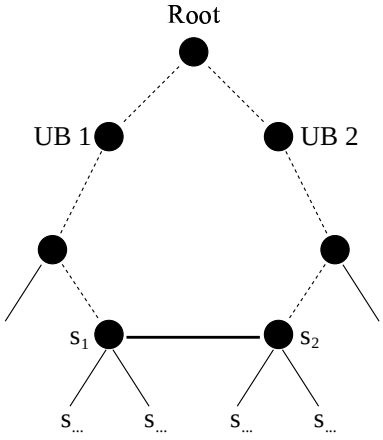


Figure 3: Local path reduction between two synsets $s_1$ and $s_2$. The dashed path is the old path, the new relation $R(s_1, s_2)$ is depicted by the thick line between $s_1$ and $s_2$.

## 6. Conclusions and future work

We have shown that the insertion of new, cross-categorial relation instances has no global impact, but a visible local impact. Words which have been far away from each other in the net now become neighbours. We expect that the linking of words of different categories will improve the usefulness of wordnets for applications such as information retrieval,

summarisation and anaphora resolution. We also expect an impact on the task of word sense disambiguation. Words might better be disambiguated by *the company they keep*.

Most of the semantic relatedness measures reported by Budanitsky and Hirst (2006) are, however, not sensitive to cross-categorial relations. We therefore see more potential in a combination of an extended wordnet with measuring semantic relatedness by random graph walks (cf. Hughes and Ramage (2007) for details).

We have currently inserted around 600 new relation instances of the verb-object types. In the near future we will investigate the impact of the new relation on the performance of semantic information retrieval and on anaphora and coreference resolution.

## 7. References


Raquel Amaro, Rui Pedro Chaves, Palmira Marrafa, and Sara Mendes. 2006. Enriching wordnets with new relations and with event and argument structures. In *Proc. CICLing 2006*, pages 28–40.

Luisa Bentivogli and Emanuele Pianta. 2003. Extending WordNet with Syntagmatic Information. In Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Second International WordNet Conference—GWC 2004*, pages 47–53, Brno, Czech Republic. Masaryk University Brno, Czech Republic.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding Dense, Weighted Connections to WordNet. In *Proceedings of the Third International WordNet Conference*, Masaryk University, Brno, Czech Republic.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using Statistics in Lexical Analysis. In Uri Zernik, editor, *Lexical acquisition: exploiting on-line resources to build a lexicon*, pages 115–164. Laurence Erlbaum Associates, Hillsdale, NJ.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic, June. Association for Computational Linguistics.

Thad Hughes and Daniel Ramage. 2007. Lexical Semantic Relatedness with Random Graph Walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 581–589, Prague, Czech Republic. Association for Computational Linguistics.

Adam Kilgarriff. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition*, pages 33–40, Falmer, Sussex.

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

Jorge Morato, Miguel Ángel Marzal, Juan Lloréns, and José Moreiro. 2003. WordNet Applications. In Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Second International WordNet Conference—GWC 2004*, pages 270–278, Brno, Czech Republic. Masaryk University Brno, Czech Republic.

Frank Henrik Müller. 2004. Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z).

Erik Tjong Kim Sang. 2007. Extracting hypernym pairs from the web. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 165–168, Prague, Czech Republic, June. Association for Computational Linguistics.

Sabine Schulte im Walde. 2006. Can Human Verb Associations help identify Salient Features for Semantic Verb Classification? In *Proceedings of the 10th Conference on Computational Natural Language Learning*, New York City, NY.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 801–808, Morristown, NJ, USA. Association for Computational Linguistics.

Eiko Yamamoto and Hitoshi Isahara. 2007. Extracting word sets with non-taxonomical relation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 141–144, Prague, Czech Republic, June. Association for Computational Linguistics.