Exploring and Navigating: Tools for GermaNet

Marc Finthammer, Irene Cramer

Faculty of Cultural Studies
Dortmund University of Technology, Germany
marc.finthammer@udo.edu, irene.cramer@udo.edu

Abstract

GermaNet is regarded to be a valuable resource for many German NLP applications, corpus research, and teaching. This demo presents three GUI-based tools meant to facilitate the exploration of and navigation through GermaNet. The GermaNet Explorer exhibits various retrieval, sort, filter and visualization functions for words/synsets and also provides an insight into the modeling of GermaNet's semantic relations as well as its representation as a graph. The GermaNet-Measure-API and GermaNet Pathfinder offer methods for the calculation of semantic relatedness based on GermaNet as a resource and the visualization of (semantic) paths between words/synsets. The GermaNet-Measure-API furthermore features a flexible interface, which facilitates the integration of all relatedness measures provided into user-defined applications. We have already used the three tools in our research on thematic chaining and thematic indexing, as a tool for the manual annotation of lexical chains, and as a resource in our courses on corpus linguistics and semantics.

1. Motivation

GermaNet (Lemnitzer and Kunze, 2002), the German equivalent of WordNet (Fellbaum, 1998), represents a valuable lexical-semantic resource for numerous German natural language processing (NLP) applications. However, in contrast to Princeton WordNet, only few graphical user interface (GUI) based tools have been created up to now for the exploration of GermaNet. In order to get an idea of it, the user is left on his own with a bunch of XML-files and insufficient means for navigation or exploration[1]. Various sub-tasks of our research in the DFG (German Research Foundation) funded project HyTex[2], such as lexical chaining or the automatic extraction of definitions, highly rely on the semantic knowledge represented in GermaNet. While intensively working with it, we accumulated a list of properties a GermaNet GUI should feature and accordingly implemented the **GermaNet Explorer**. In addition, during the course of our research on lexical chaining for German corpora (Cramer and Finthammer, 2008), we also investigated semantic relatedness and similarity measures based on GermaNet as a resource. The results of this work led us to the implementation of eight GermaNet-based relatedness measures, which we provide as Java-API, the so-called **GermaNet-Measure-API**. In order to facilitate the use, we also developed a GUI for this API, the so-called **GermaNet Pathfinder**. We think that these three tools simplify the use of and work with GermaNet, since they can be integrated into various NLP applications and can also be used as a resource for the visual exploration of and navigation through GermaNet. All three tools are freely available for download.

[1] As a matter of fact, relatively few computational linguists and language researchers work with German data, consequently, the development of tools for German resources, such as GermaNet, takes more time.

[2] The HyTex project aims at the development of text-to-hypertext conversion strategies based on text-grammatical features. Please, refer to our project web pages http://www.hytex.info/ for more information about our work.

2. GermaNet Explorer

Many researchers working with GermaNet have the same experience: they lose their way in the rich, complex structure of its XML-representation. In order to solve this problem, we implemented the GermaNet Explorer, of which a screenshot is shown in Figure 1. Its most important features are: the word sense retrieval function (Figure 1, region 1) and the structured presentation of all semantic relations pointing to/from the synonym set (synset) containing the currently selected word sense (Figure 1, region 2). In addition, the GermaNet Explorer offers a visual, graph-based navigation function. A synset (in Figure 2 [`Rasen, Grünfläche`] Engl. lawn) is displayed in the center of a navigation graph surrounded by its direct semantically related synsets, such as hypernyms (in Figure 2 [`Nutzfläche, Grünland`]) above the current synset, hyponyms (in Figure 2 [`Kunstrasen, Kunststoffrasen`] and [`Grüngürtel`]) below, holonyms (in Figure 2 [`Grünanlage, Gartenanlage, Eremitage`]) to the left, and meronyms (in Figure 2 [`Graspflanze, Gras`]) to the right. In order to navigate the graph representation of GermaNet, one simply clicks on a related synset, in other words one of the rectangles surrounding the current synset shown in Figure 2. Subsequently, the visualization is refreshed: the selected synset moves into the center of the displayed graph and the semantically related synsets are updated accordingly.

In addition, the GermaNet Explorer features a representation of all synsets, which is illustrated in Figure 3, region 1. It also provides retrieval, filter, and sort functions (Figure 3, region 2). Moreover, the GermaNet Explorer exhibits the same function as shown in Figure 3 and a similar GUI for the list of all word senses. We found that these functions, both for the word senses and the synsets, provide a very detailed insight into the modeling and structure of GermaNet and thus helped us to understand its strengths and weaknesses.

We were already able to successfully utilize the GermaNet Explorer in various areas of our research and teaching. E.g. in experiments on the manual annotation of lexical chains in

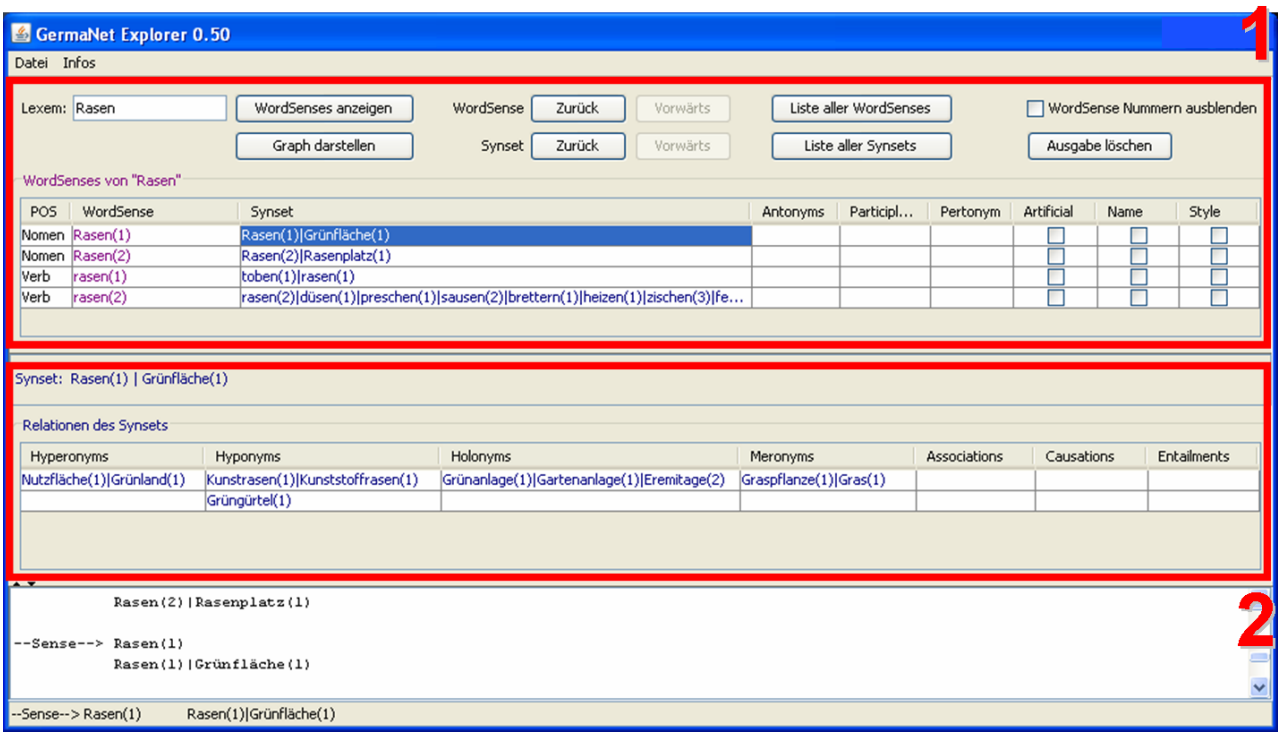

Figure 1: Screenshot GermaNet Explorer

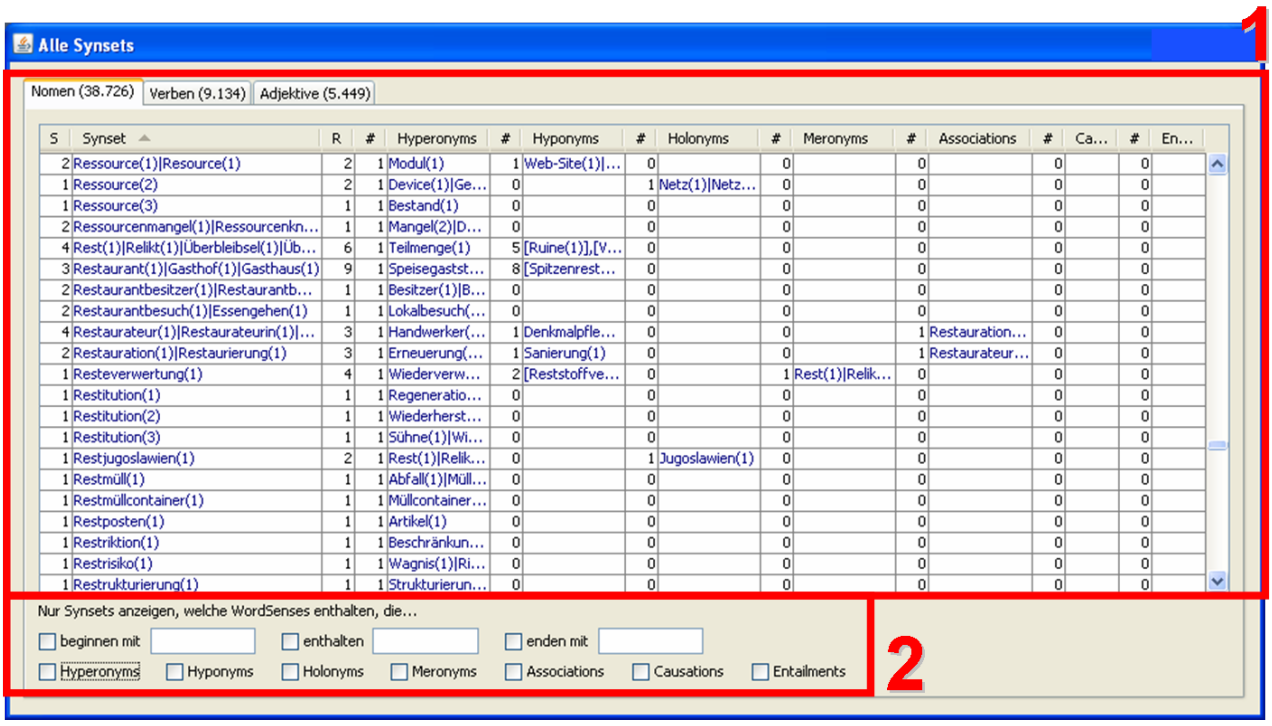

Figure 3: Screenshot GermaNet Explorer – Representation of the List of All GermaNet Synsets

German corpora, our subjects used the GermaNet Explorer to find paths representing semantic relatedness between two words. This work is partially described in (Stührenberg et al., 2007) and (Cramer et al., to appear). We also found it helpful for the visualization of lexical semantic concepts and thus for the training of our students in courses on e.g. semantics or corpus linguistics. We hence argue that the GermaNet Explorer represents a tool which is applicable in many scenarios.

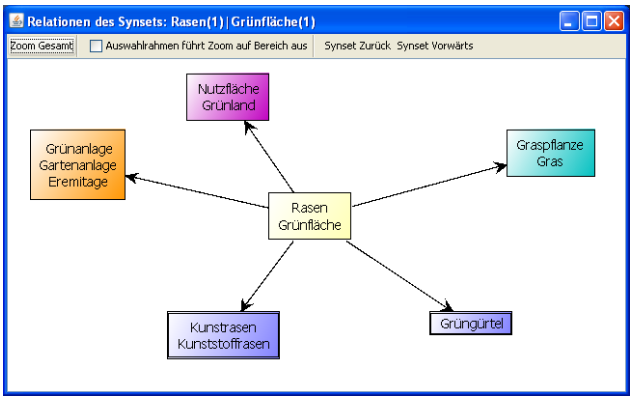

Figure 2: Screenshot GermaNet Explorer – Visual Graph Representation

3. GermaNet Measure-API and Pathfinder

Semantic relatedness measures express *how much two words have to do with each other*. This represents an essential information in various NLP applications and is extensively discussed in the literature e.g. (Budanitsky and Hirst, 2001). Many measures have already been investigated and implemented for Princeton WordNet, however, there are only few publications addressing measures based on GermaNet e.g. (Gurevych and Niederlich, 2005) and (Cramer and Finthammer, 2008). The calculation of semantic relatedness is a subtask of our research in HyTex. We implemented eight GermaNet based measures[3] e.g. (Jiang and Conrath, 1997), (Leacock and Chodorow, 1998), (Lin, 1998), (Resnik, 1995), (Wu and Palmer, 1994) and three Google[4] based measures e.g. (Cilibrasi and Vitanyi, 2007). Because of the compared to Princeton WordNet different structure of GermaNet, it was necessary to adapt the algorithms discussed in the literature and in parts already available for Princeton WordNet.

The GermaNet-Measure-API consists of a hierarchically organized collection of measure classes, all of these (GermaNet or Google based) exhibit the same interface, which provides methods for many elementary operations, e.g. for a given pair of words (or synsets) the calculation of the relatedness. Some of the measures such as (Jiang and Conrath, 1997) are defined as a distance measure, therefore the GermaNet-Measure-API provides conversion methods to map between distance and relatedness measure values. The above mentioned measure interface also facilitates the integration of all measures into any given application. Moreover, the measures can be easily exchanged in order to explore the impact on the performance of a specific application. Further, for a given set of word pairs the values of all measures can be computed simultaneously, which enables the user to compare the different measures with respect to a specific research objective.

However, the interpretation of a specific measure value is not a trivial task, as most measures are non-linear and continuous, and since moreover many applications do not depend on continuous measure values the GermaNet-Measure-API also provides a mapping into the three discrete classes: `not related`, `related`, and `strongly related`[5].

In order to additionally facilitate the integration of all measures into user-defined applications and to allow the straightforward comparison and evaluation of the different measures, we also implemented a GUI, the GermaNet Pathfinder, shown in Figure 4. It provides a user-friendly access to all functions of the GermaNet-Measure-API. The GermaNet Pathfinder exhibits the following features:

- Selection of one / all available measures to calculate the relatedness for a given word pair or synset pair (Figure 4, region 1).
- Calculation of the relatedness with respect to all possible readings–if applicable all word senses and thus synset combinations are considered (Figure 4, region 2).
- Mapping between relatedness value and relatedness class.
- Visualization of the respective path between two synsets (Figure 5).
- Automatic calculation of the relatedness (or respective relatedness class) for a complete list of word pairs (provided as a text file in .csv-format).
- Storing of the results as .csv-file for evaluation with an external application.
- Inspection and extraction of many features of GermaNet and its hyponym-tree representation. (GermaNet Pathfinder provides direct access to all of the involved hyponym-tree features, e.g. the shortest path between two synsets, the least common subsumer of two synsets, the parents, children, siblings, descendants of a synset, the minimum and maximum depth of a synset, and the information content of a synset.)
- Visualization of synsets and subtrees (i.e. the subtree with a given synset as root) as well as the complete hyponym-tree. (This representation shows all descendents of the selected synset and therefore provides a comprehensive overview of a given part of GermaNet, see Figure 4, region 3).

The GermaNet-Measure-API has already been successfully used in our lexical chainer, called `GLexi`, see (Cramer and Finthammer, 2008) and (Cramer et al., to appear) for further information. We also found the GermaNet Pathfinder very helpful to explore GermaNet and retrace semantically motivated paths, which is illustrated in Figure 5 as the (shortest)

[3]For more information about the measures implemented as well as our research on lexical/thematic chaining and the performance of our GermaNet based lexical chainer, please refer to (Cramer and Finthammer, 2008) and (Cramer et al., to appear).

[4]The three Google measures are based on co-occurence counts and realize different algorithms to convert these counts into values representing semantic relatedness.

[5]The thresholds for this mapping were determined on the basis of human-judgement experiments partially described in (Cramer and Finthammer, 2008).

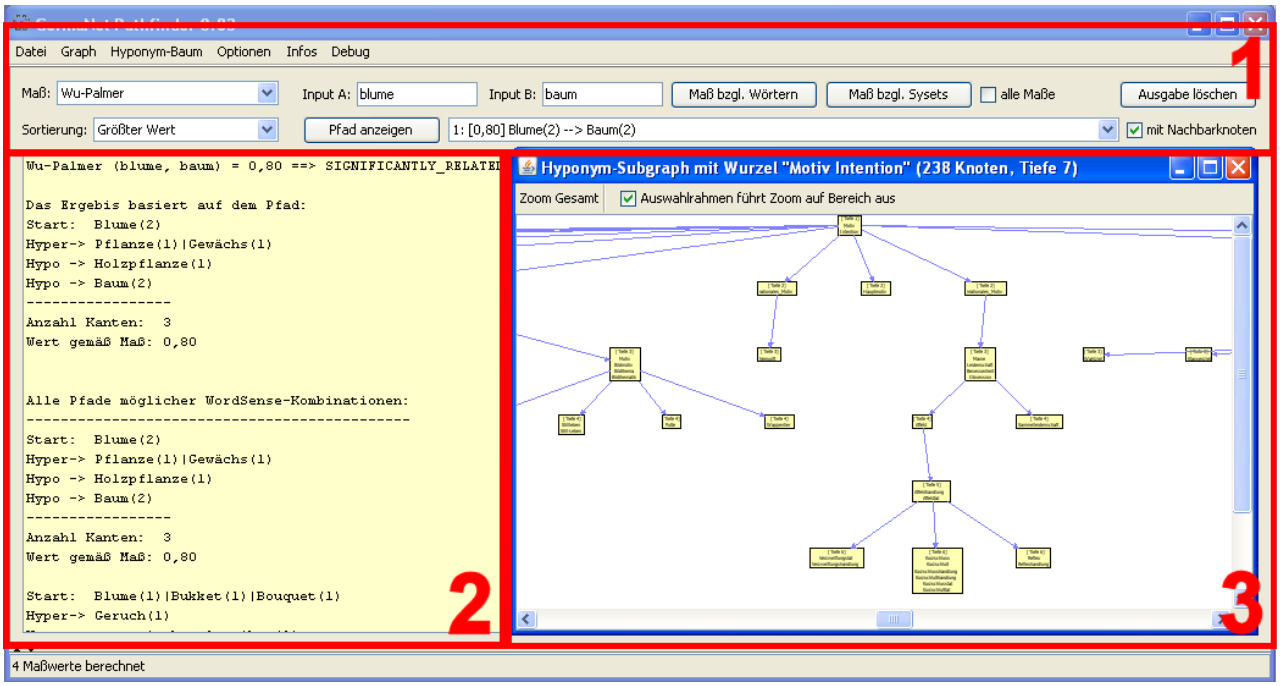

Figure 4: Screenshot GermaNet Pathfinder

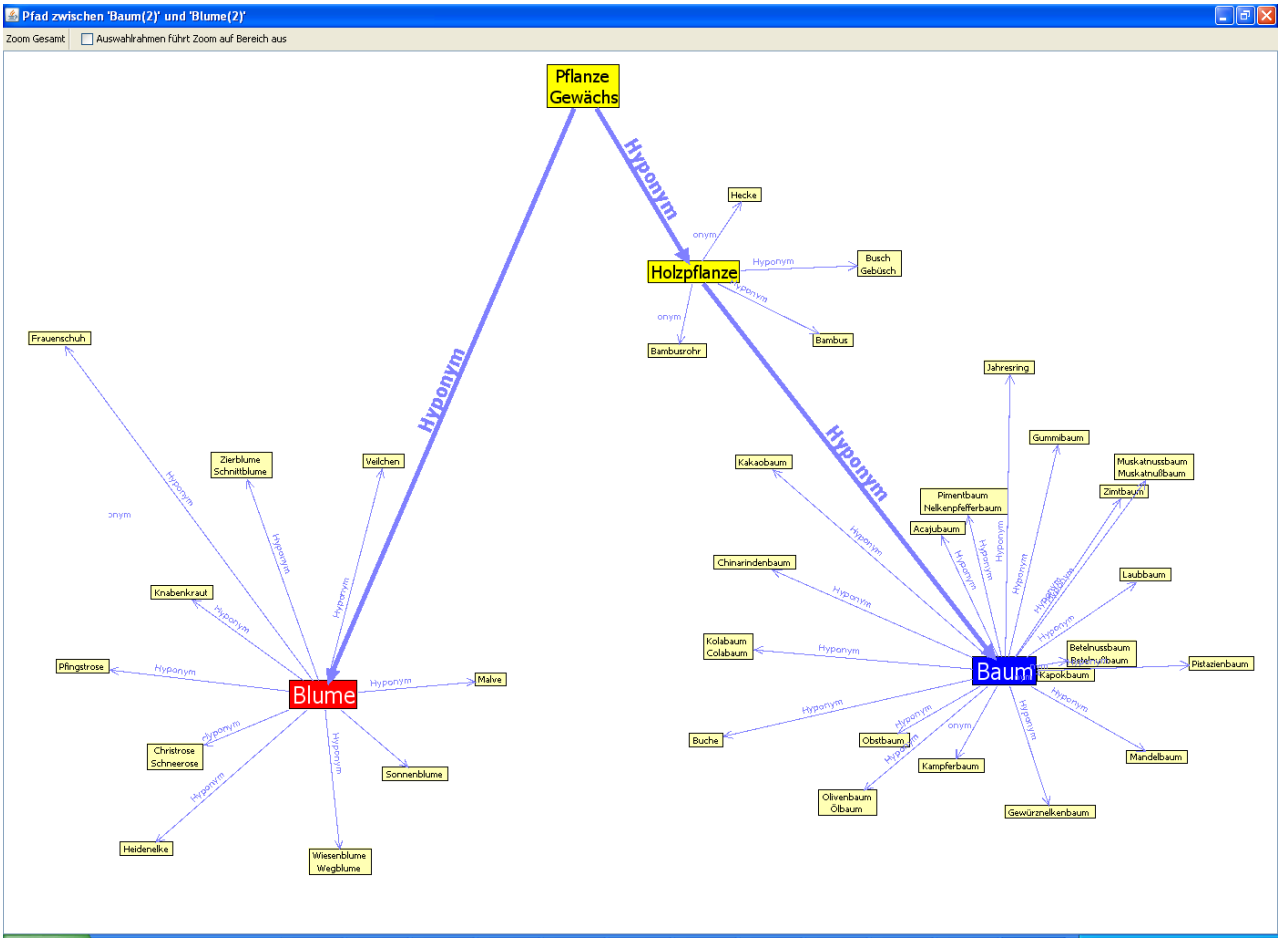

Figure 5: Screenshot GermaNet Pathfinder – Illustration of a Shortest Path Between `Blume` and `Baum`

path between `Blume` (Engl. flower) and `Baum` (Engl. tree). This path consists of three steps (hypernymy – hyponymy – hyponymy) and traverses two synsets; it thus represents the kind of (indirect) semantic relation relevant in e.g. lexical chaining.

4. Open Issues and Future Work

We have already used the GermaNet Explorer, GermaNet-Measure-API and Pathfinder in our research on thematic chaining, as a tool for the manual annotation of lexical chains and as a resource in our seminars. In our future work, we plan to further explore the possible fields of application, e.g. for training students and annotators. The research on relatedness measures both for GermaNet and Princeton WordNet among others (Budanitsky and Hirst, 2001) shows that the established algorithms are not yet able to satisfactorily represent the semantic relations between two words. Namely, human-judgement experiments stress that the correlation between the relatedness measures and the intuition of subjects is much to low. We therefore plan to investigate alternative relatedness measures, which we also intend to integrate into the GermaNet Pathfinder. However, the usefulness of the GermaNet Explorer and Pathfinder is constrained by the coverage and modeling quality of the underlying semantic lexicon. Therefore, we also hope to hereby provide tools to see behind GermaNet's curtain and to thus facilitate the user-centered work with this interesting and valuable resource.

5. References

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources at NAACL-2000*.

Rudi Cilibrasi and Paul M. B. Vitanyi. 2007. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19.

Irene Cramer and Marc Finthammer. 2008. An evaluation procedure forword net based lexical chaining: Methods and issues. In *Proc. of the 4th Global Wordnet Conference, Szeged, Hungary, 2008*.

Irene Cramer, Marc Finthammer, Alexander Kurek, Lukas Sowa, Melina Wachtling, Tobias Claas, and Angelika Storrer. to appear. Experiments on lexical chaining for german corpora: Annotation, extraction, and application. In *Ontologies and Semantic Lexica in Automated Discourse Analysis*. LDV-Forum, GLDV.

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Iryna Gurevych and Hendrik Niederlich. 2005. Computing semantic relatedness in german with revised information content metrics. In *Proc. of OntoLex 2005 - Ontologies and Lexical Resources, IJCNLP 05 Workshop*.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Proc. of the International Conference on Research in Computational Linguisics*.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*.

Lothar Lemnitzer and Claudia Kunze. 2002. Germanet - representation, visualization, application. In *Proc. of the Language Resources and Evaluation Conference (LREC2002)*.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proc. of the 15th International Conference on Machine Learning*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the IJCAI 1995*.

Maik Stührenberg, Daniela Goecke, Nils Diewald, Alexander Mehler, and Irene Cramer. 2007. Web-based annotation of anaphoric relations and lexical chains. In *Proc. of the Linguistic Annotation Workshop, ACL 2007*.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*.