Tools for Exploring GermaNet in the Context of CL-Teaching

Irene Cramer and Marc Finthammer

Abstract. Word nets, such as Princeton WordNet or GermaNet, are resources organizing a (more or less extensive) fraction of the vocabulary of a language according to lexical semantic relations. Such resources are widely used in natural language processing (NLP) and computational linguistics (CL) both for research and teaching purposes. While several graphical user interfaces (GUI) exist for Princeton WordNet–some of which are also available online –GermaNet still lacks such utilities. In this paper we describe two GUI-based tools meant to facilitate the navigation through and exploration of GermaNet. Both are freely available for download from our project web page (`www.hytex.info`). We additionally discuss ways of deploying these tools in teaching. We argue that the GUI-based access to word nets, which can be regarded as a fundamental resource in CL and NLP, enhances the students' understanding of basic lexical semantic concepts, computational semantics and lexicography.

1 Motivation

Word nets are lexical semantic resources modeled according to the principles introduced in Princeton WordNet (e.g. Fellbaum 1998). The central idea of word nets is to group synonymous lexical units, also including compounds and multi-word-units, into so-called synsets (synonym sets) and link them according to lexical semantic relations, such as hyponymy, meronymy, antonymy etc. Currently, Princeton WordNet (Version 3.0) contains approximately 150,000 synsets[1] and approximately 200,000 lexical units. The conceptual design and the resource itself are upgraded continuously–e.g. over the past years proper names have been added and tagged accordingly (Miller and Hristea 2006) and non-classical, i.e. psycho-linguistically motivated, link types have been included as an additional layer of relations (Boyd-Graber et al. 2006). Many NLP-applications, such as information retrieval and information extraction (e.g. Mandala et al. 1998) or word sense disambiguation (e.g. Banerjee and Pedersen 2002), highly rely on word nets as a (lexical) semantic resource [2]. Therefore, in recent years, word nets have been developed for many languages, e.g. in the context of EuroWordNet (Vossen 1998) for seven European lan-

1. Please refer to `http://wordnet.princeton.edu/man/wnstats.7WN` for more information.
2. Cp. Fellbaum (1998), Kunze (2001), and the Proceedings of the Global WordNet Conferences e.g. Tanács et al. (2008)

guages, and connected via the so-called ILI[3]. GermaNet, the German counterpart of Princeton WordNet, which has been developed since 1997 at the University of Tübingen, currently (Version 5.1) consists of 58,000 synsets and 82,000 lexical units[4].

As word nets constitute a fundamental resource in many NLP-applications, they should also play a major role in CL curricula and be carefully introduced in courses on e.g. computational semantics and NLP resources. In addition to the modeling and structure of word nets, students should be familiarized with algorithms for the calculation of semantic relatedness, similarity, and distance (cp. Budanitsky and Hirst 2006; Patwardhan and Pedersen 2006) –both from a theoretical and a practical point of view. Such algorithms are regarded as a fundamental component in various NLP-applications, such as text summarization (e.g. Barzilay and Elhadad 1997), malapropism recognition (e.g. Hirst and St-Onge 1998), automatic hyperlink generation (e.g. Green 1999), question answering (e.g. Novischi and Moldovan 2006), and topic detection/topic tracking (e.g. Carthy 2004). And even for traditional courses on e.g. semantics, word nets offer interesting options. Typically, semantic relations are introduced providing a few more or less plausible examples (e.g. `Rappe`, Engl. black horse, is a hyponym of horse). In contrast, Princeton WordNet[5] and GermaNet offer plenty of illustrative material, since they both cover a wide range of lexical units connected via semantic relations. While Princeton WordNet already exhibits several GUI-based interfaces, some of which are also available online [6], GermaNet still lacks such utilities. This might have two causes: firstly, the research community working with German data is much smaller than the one working with English; secondly, some word nets are subject to particular license restrictions[7]. In addition, GermaNet differs form Princeton WordNet with respect to some modeling aspects; therefore, tools implemented for WordNet cannot be adopted for GermaNet in its current state. While implementing a lexical chainer–called GLexi, cp. Cramer and Finthammer (2008)–for German specialized domain corpora, Finthammer and Cramer (2008) implemented two GUI-based tools for the exploration of GermaNet.

Sections 2 and 3 introduce these tools and their basic features. Most researchers working with GermaNet share the same experience of getting lost in the rich structure of its XML-representation. Thus, the GUI-based tools implemented by Finthammer and Cramer (2008) are meant to help both researchers and students explore

3. ILI stands for interlingual index. Please refer to `http://www.illc.uva.nl/EuroWordNet/` for more information.
4. Please refer to `http://www.sfs.uni-tuebingen.de/lsd/` for more information.
5. Princeton WordNet also features glosses explaining the meaning of a lexical unit and example sentences; it thus represents a full-fledged digital dictionary, which could be used in various application scenarios, e.g. as an interesting and innovative resource in classes of (computational) lexicography.
6. E.g. WordNet Browser (`http://wordnet.princeton.edu/perl/webwn`) or Vocabulary Helper (`http://poets.notredame.ac.jp/cgi-bin/wn`).
7. Please refer to `http://www.sfs.uni-tuebingen.de/lsd/` for more information on this issue.

GermaNet[8]. In this paper, we also discuss possibilities of how to utilize the tools in CL courses, especially practical sessions on lexical/computational semantics or computational lexicography. We have already used the tools in an annotation experiment (Cramer et al. accepted) with first-year and second-year students. We found that students employing the two GUI-based tools need less training than the students employing the XML-representation of GermaNet only. We also think that the GUI-based GermaNet interfaces might enhance the students' understanding of basic lexical semantic concepts. We therefore sketch some ideas of practical sessions introducing GermaNet and semantic relatedness measures drawing on the two tools in the following sections.

2 GermaNet Explorer

GermaNet Explorer, of which a screenshot is shown in Figure 1, is a tool for exploration and retrieval. Its most important features are: the word sense retrieval function (Figure 2) and the structured presentation of all semantic relations pointing to/from the synset containing the currently selected word sense (Figure 3). The GermaNet

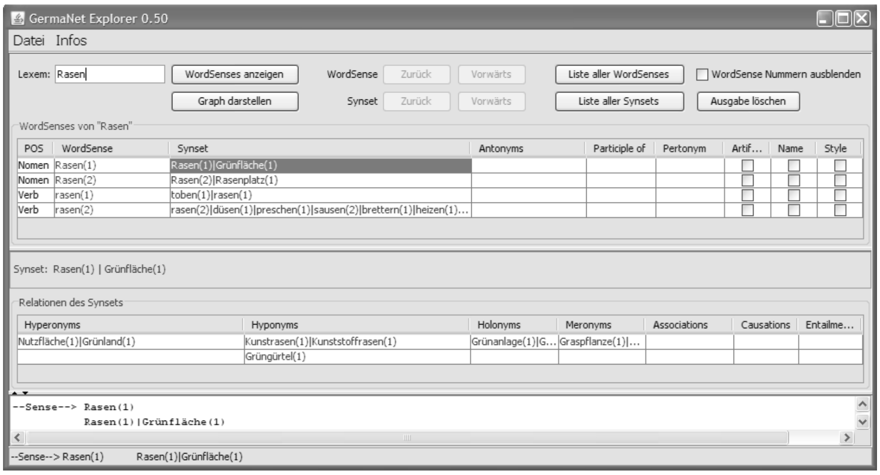


Figure 1. Screenshot GermaNet Explorer

Explorer also provides a visual, graph-based navigation function: a synset (in Figure 4 `[Rasen, Grünfläche]` Engl. lawn) is displayed in the center of a navigation

8. The tools have been implemented in the context of the DFG-funded project HyTex and are freely available for download from our project web page (`www.hytex.info`).

graph surrounded by its direct semantically related synsets, such as hypernyms (in Figure 4 [`Nutzfläche, Grünland`]) above the current synset, hyponyms (in Figure 4 [`Kunstrasen, Kunststoffrasen`] and [`Grüngürtel`]) below, holonyms (in Figure 4 [`Grünanlage, Gartenanlage, Eremitage`]) to the left, and meronyms (in Figure 4 [`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 4. Subsequently, the

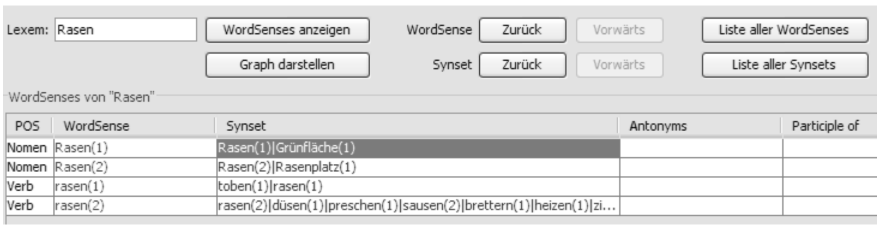


Figure 2. Screenshot GermaNet Explorer: Retrieval Functions

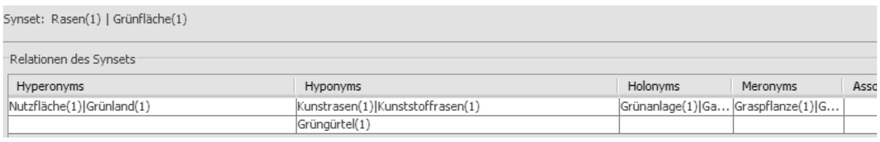


Figure 3. Screenshot GermaNet Explorer: Relations Pointing to/from Current Synset

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 5. It also provides retrieval, filter, and sort functions (Figure 6). Moreover, the GermaNet Explorer exhibits the same functions as shown in Figures 5 and 6 with 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. E.g. in a hands-on session of a (computational) semantics course, using the GermaNet Explorer students can (visually) examine lexical semantic relations for a (relatively) large fraction of the German vocabulary. While exploring sub-sets of lexical units, they can also compare their own intuition as well as the intuition of a group of German native speakers (namely, their fellow students) about the semantic relations between these lexical units with the modeling present in GermaNet. Potentially observed differences between their own intuition, the intuition of their fellow students, and GermaNet will certainly raise their awareness of lexical semantic concepts and the challenge of building such a resource consistently.

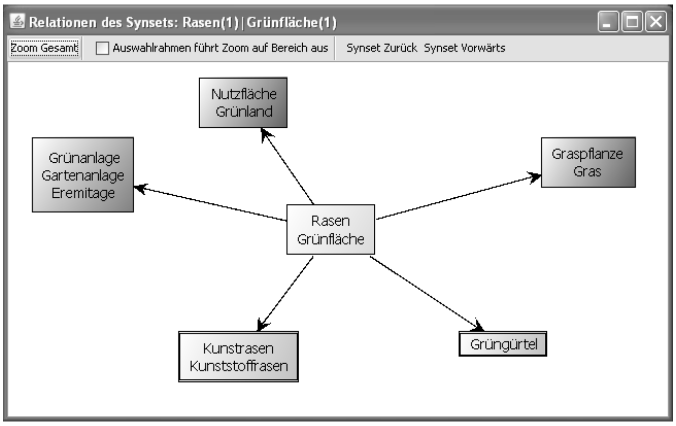
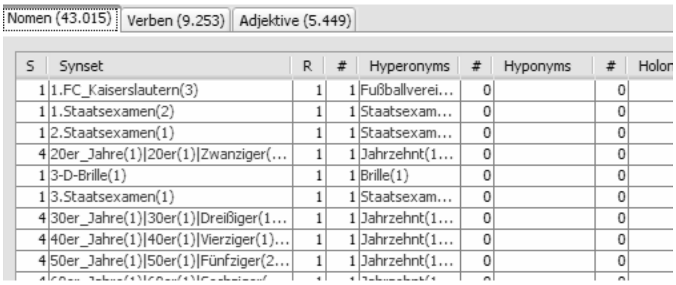


Figure 4. Screenshot GermaNet Explorer: Visual Graph Representation



Nomen (43.015) | Verben (9.253) | Adjektive (5.449)

S	Synset	R	#	Hyperonyms	#	Hyponyms	#	Holon
1	1.FC_Kaiserslautern(3)	1	1	Fußballverei...	0		0	
1	1.Staatsexamen(2)	1	1	Staatsexam...	0		0	
1	2.Staatsexamen(1)	1	1	Staatsexam...	0		0	
4	20er_Jahre(1)|20er(1)|Zwanziger(...	1	1	Jahrzehnt(1...	0		0	
1	3-D-Brille(1)	1	1	Brille(1)	0		0	
1	3.Staatsexamen(1)	1	1	Staatsexam...	0		0	
4	30er_Jahre(1)|30er(1)|Dreißiger(1...	1	1	Jahrzehnt(1...	0		0	
4	40er_Jahre(1)|40er(1)|Vierziger(1)...	1	1	Jahrzehnt(1...	0		0	
4	50er_Jahre(1)|50er(1)|Fünfziger(2...	1	1	Jahrzehnt(1...	0		0	

Figure 5. Screenshot GermaNet Explorer: List of All GermaNet Synsets

Last but not least, simple corpus-based methods to extract semantic relations (such as the well-known Hearst patterns, cp. Hearst (1992)) may be compared with relations in GermaNet. An example is shown in Figure 7. Moreover, by contrasting relations of the same type, the students can learn to discern differences in relation strength and semantic distance (cp. Boyd-Graber et al. (2006))[9]. Examples are shown in Figures 8 and 9. Obviously, the modeling of the synsets containing terminology, such as *bumble-bee*, is much more fine-grained than the one of synsets containing general concepts, such as *money*. The synsets of `fliegen` (Engl. to fly) and `kauen` (Engl.

9. Boyd-Graber et al. (2006) cite the following as an example: "It is intuitively clear that the semantic distance between the members of hierarchically related pairs is not always the same. Thus, the synset [run] is a subordinate of [move], and [jog] is a subordinate of [run]. But [run] and [jog] are semantically much closer than [run] and [move]."

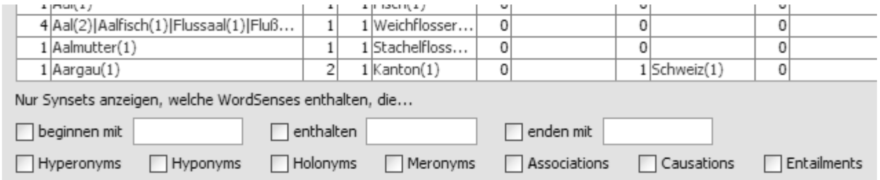


Figure 6. Screenshot GermaNet Explorer: List of All GermaNet Synsets: Filter and Search Functions

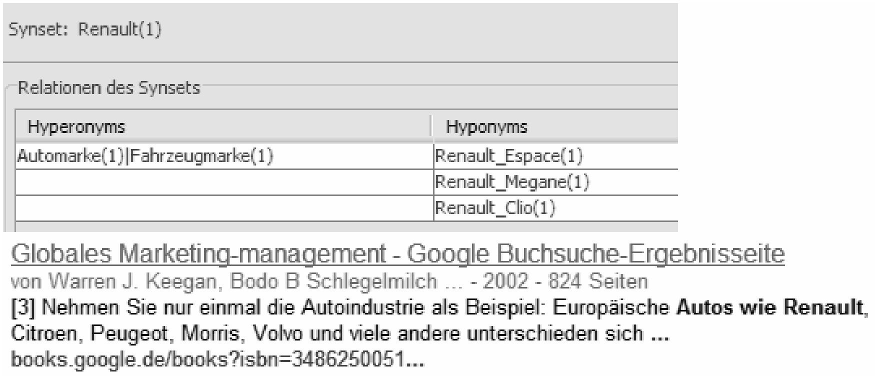


Figure 7. Hyponym Relation: GermaNet vs. Pattern-Based Corpus Approach

to chew) are both directly connected with `schwingen/oszillieren` (Engl. to oscillate), which implies that it only takes two steps from *fly* to *chew*. We assume that these and similar examples can improve the students' understanding of the structure and potential shortcomings of word nets in general and GermaNet in particular. Finally, even the often criticized lack of glosses in GermaNet may be used productively in order to discuss (sometimes subtle) differences in the meaning of lexical units or synsets. I.e. the meaning of synsets and lexical units can be retraced on the basis of the lexical semantic relations modeled in GermaNet. As an experiment, the thus extracted information can again be exploited by the students to write glosses and example sentences.

3 GermaNet Pathfinder

As mentioned in Section 1, the calculation of semantic relatedness, similarity, and distance plays a crucial role in many NLP-applications. Those measures express how much two words have to do with each other; they are extensively discussed in the

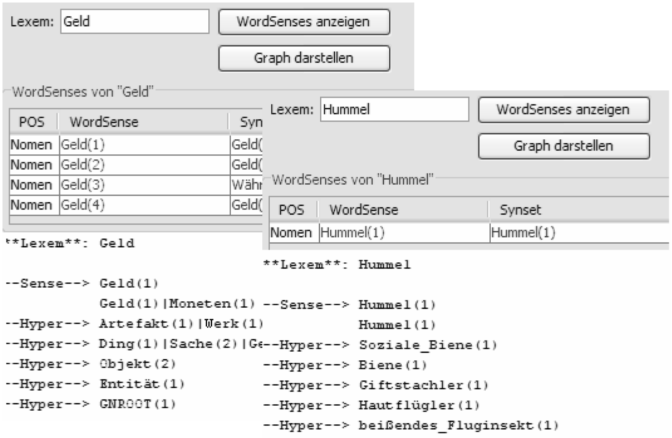


Figure 8. Coarse-Grained vs. Fine-Grained Modeling of Synsets

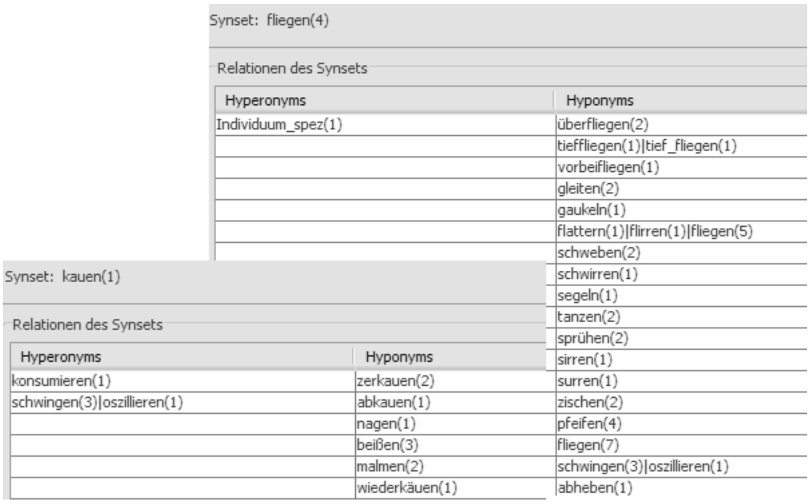


Figure 9. Two Steps form *Fly* to *Chew* Via *Oscillate*

literature (e.g. Budanitsky and Hirst 2006). Many measures have already been investigated and implemented for Princeton WordNet (e.g. Patwardhan and Pedersen 2006), however, there are only a few publications addressing measures based on GermaNet (e.g. Finthammer and Cramer 2008; Gurevych and Niederlich 2005). The GermaNet Pathfinder constitutes a GUI-based tool which has been developed as a central component of the lexical chainer GLexi (Cramer and Finthammer 2008). It implements eleven semantic measures–eight GermaNet-based[10] and three Google-based[11] ones–and integrates all measures into a common Java-API. The GermaNet Pathfinder additionally features a GUI meant to facilitate the intellectual analysis of semantic distance between given pairs of synsets or lexical units with respect to one semantic measure, a subset, or all. In short, the GermaNet Pathfinder exhibits the

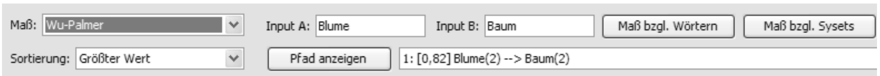


Figure 10. Screenshot GermaNet Pathfinder

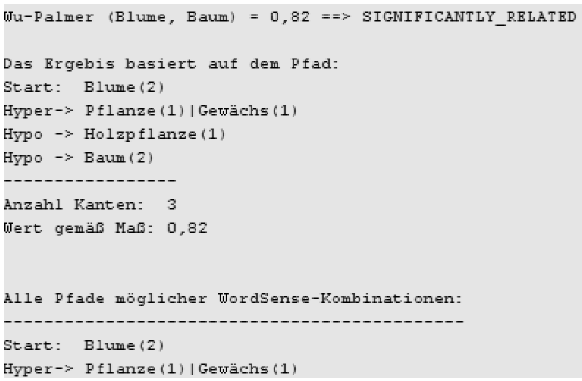


Figure 11. Screenshot GermaNet Pathfinder: Shortest Path and All Possible Paths

following features: In order to calculate the relatedness for a given word-pair or pair of synsets (see Figure 10), the user may select a single measure or all measures at one time. Furthermore, the relatedness values can be calculated with respect to all possible synset-synset combinations or one particular combination (see Figure 11). In order to analyze and compare the relatedness values of the different measures, the

10. For more information on the measures implemented as well as the research on lexical/thematic chaining and the performance of GLexi, the lexical chainer for German corpora, please refer to Cramer and Finthammer (2008) and Cramer et al. (accepted) respectively. See e.g. Jiang and Conrath (1997), Leacock and Chodorow (1998), Lin (1998), Resnik (1995), Wu and Palmer (1994) for more information on semantic measures drawing on word nets.
11. The three Google measures are based on co-occurrence counts and use different algorithms to convert these counts into values representing semantic relatedness. See e.g. Cilibrasi and Vitanyi (2007) for more information on this.

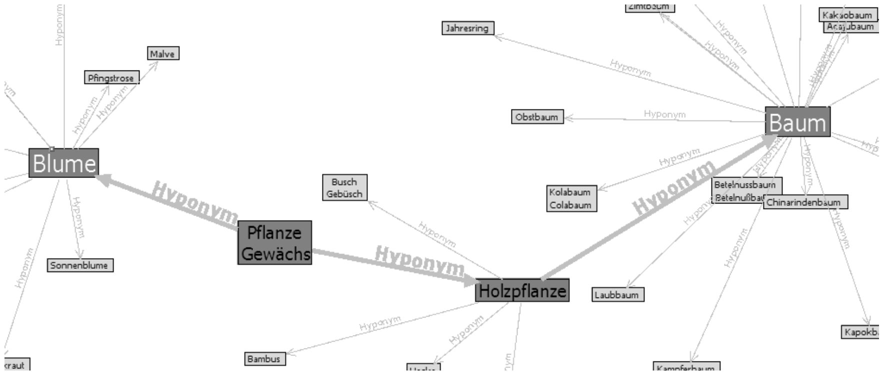


Figure 12. Screenshot GermaNet Pathfinder: Path Visualization Function

GermaNet Pathfinder includes a function to calculate the relatedness for a complete list of word-pairs or pairs of synsets, the results of which are stored in .cvs-format. We found that this function is especially useful for evaluating the performance of semantic measures with regard to a given task. Finally, the relatedness value, which corresponds–in the case of the GermaNet-based measures–to a path, can be examined visually using the corresponding GermaNet Pathfinder functions (see Figure 12).

Semantic relatedness measures play a key role in NLP and consequently represent central components of many applications. Therefore, CL students should be familiar with the basic concepts and algorithms of semantic relatedness. However, in order to understand and be able to independently and productively use semantic relatedness measures, theoretical and practical knowledge (or experience) is required. For this purpose a hands-on seminar might be most suitable. Using the Pathfinder, CL students may explore the various aspects of GermaNet considered in the calculation of semantic relatedness, such as path length, relation types, and graph depth. By comparing various paths between a given word pair, subtle differences between the algorithms can be discussed. An example is shown in Figure 13, which illustrates the difference between relatedness measures drawing on the complete GermaNet graph as a resource and those exclusively using the hyponym-tree. The GermaNet Explorer and Pathfinder may also help analytically retrace paths between semantically more or less related pairs of words or synsets. These manually constructed paths (which in this case correspond to semantic relatedness values) can then be compared with the automatically calculated ones. This might raise the students' awareness of aspects in need of improvement: on the one hand in the modeling of GermaNet and on the other hand in the algorithms of the semantic relatedness measures. The example shown in Figure 13 (i.e. the path between `Kuh`, Engl. cow, and `Milch`, Engl.

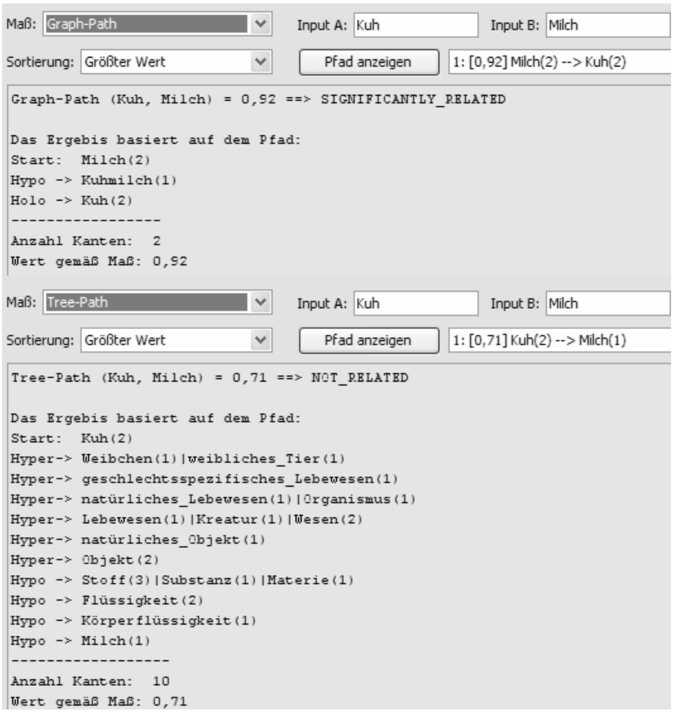


Figure 13. Path Based on Complete GermaNet-Graph (Graph-Path) vs. Path Based on Hyponym-Tree (Tree-Path)

milk) demonstrates that some (if not most) paths do not reproduce human intuition and thus differ from the paths humans would select. Finally, the comparison of the three Google-based measures (relying on corpus statistics) and the eight GermaNet-based ones (relying on manually created structures) may clarify which aspects of the human intuition on semantic distance are included in the measures using different resources. Further, the comparison may highlight the ways in which the measures diverge, i.e. syntagmatic vs. paradigmatic relations or simply coverage.

4 Outlook

We plan to continue introducing word nets, lexical semantic concepts, and algorithms of semantic relatedness using the GermaNet Explorer and GermaNet Pathfinder. As mentioned above, we found that the tools might indeed support the learning process of our students. However, we think in order to successfully employ the two tools it

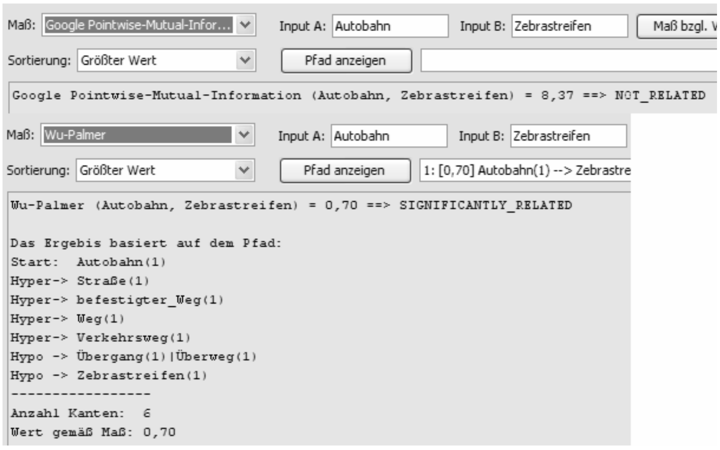


Figure 14. Google-based vs. GermaNet-based Measure: Freeway – Crosswalk

will be necessary to carefully design, deploy, and evaluate seminar sessions. Therefore, we plan to employ both in our courses in an even more focused manner. In doing so, we intend to collect information on the following aspects:

- How may lessons on lexical semantics be enriched by using both tools?
- How may word nets, such as Princeton WordNet or GermaNet, be effectively introduced and presented?
- Which tasks or student projects respectively are suitable for accomplishing this objective?

In recent years, academic higher education teachers (particularly, but not only in computational linguistics) have dedicated a considerable amount of commitment to the development of courses. The exchange of ideas at conferences and workshops devoted to this topic[12] has disclosed many interesting experiences. We argue that it is worthwhile to write these up so that more teachers may benefit from these insights. Consquently, we plan to test our ideas for tasks as outlined in Sections 2 and 3 in our courses. When the first positive experiences can be validated, we intend to compile a teaching plan and make it publicly available[13].

12. See e.g. TeachCL-08 (`http://verbs.colorado.edu/teachCL-08/`)
13. We will possibly also publish it on the ACL wiki, which provides a repository of teaching material (`http://aclweb.org/aclwiki/index.php?title=Teaching`).

References

Banerjee, Satanjeev and Ted Pedersen (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, 136–145, Springer.

Barzilay, Regina and Michael Elhadad (1997). Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop*, 10–17.

Boyd-Graber, J., C. Fellbaum, D. Osherson, and R. Schapire (2006). Adding dense, weighted, connections to wordnet. In *Proceedings of the 3rd Global WordNet Meeting*, 29–35.

Budanitsky, Alexander and Graeme Hirst (2006). Evaluating wordnet-based measures of semantic relatedness. *Computational Linguistics* 32 (1):13–47.

Carthy, Joe (2004). Lexical chains versus keywords for topic tracking. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, 507–510, Springer.

Cilibrasi, Rudi and Paul M. B. Vitanyi (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering* 19(3):370–383.

Cramer, Irene and Marc Finthammer (2008). An evaluation procedure for word net based lexical chaining: Methods and issues. In *Proceedings of the 4th Global WordNet Meeting*, 120–147.

Cramer, Irene, Marc Finthammer, Alexander Kurek, Lukas Sowa, Melina Wachtling, and Tobias Claas (accepted). Experiments on lexical chaining for german corpora: Annotation, extraction, and application. *LDV-Forum* Ontologies and Semantic Lexical in Automated Discourse Analysis.

Fellbaum, Christiane (ed.) (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Finthammer, Marc and Irene Cramer (2008). Exploring and navigating: Tools for germanet. In *Proceedings of the 6th Language Resources and Evaluation Conference*.

Green, Stephen J. (1999). Building hypertext links by computing semantic similarity. *IEEE Transactions on Knowledge and Data Engineering* 11(5):713–730.

Gurevych, Iryna and Hendrik Niederlich (2005). Accessing germanet data and computing semantic relatedness. In *Companion Volume of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005)*, 5–8.

Hearst, Marti A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, 539–545.

Hirst, Graeme and David St-Onge (1998). Lexical chains as representation of context for the detection and correction malapropisms. In Christiane Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, 305–332, The MIT Press.

Jiang, Jay J. and David W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of ROCLING X*, 19–33.

Kunze, Claudia (2001). *Computerlinguistik und Sprachtechnologie: Eine Einführung*, chapter Lexikalisch-semantische Wortnetze, 386–393. Spektrum.

Leacock, Claudia and Martin Chodorow (1998). Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, 265–284, The MIT Press.

Lin, Dekang (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, 296–304.

Mandala, Rila, Takenobu Tokunaga, and Hozumi Tanaka (1998). The use of WordNet in information retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 31–37.

Miller, George A. and Florentina Hristea (2006). Wordnet nouns: Classes and instances. *Computational Linguistics* 32(1):1–3.

Novischi, Adrian and Dan Moldovan (2006). Question answering with lexical chains propagating verb arguments. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 897–904.

Patwardhan, Siddharth and Ted Pedersen (2006). Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *EACL Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, 1–8.

Resnik, Philip (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the IJCAI 1995*, 448–453.

Tanács, Attila, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen (eds.) (2008). *Proceedings of the 4th Global WordNet Conference*, University of Szeged, Department of Informatics.

Vossen, Piek (ed.) (1998). *EuroWordNet: a multilingual database with lexical semantic networks*, Kluwer Academic Publishers.

Wu, Zhibiao and Martha Palmer (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 133–138.