

Experiments on Lexical Chaining for German Corpora: Annotation, Extraction, and Application

1 Motivation

Converting linear text documents into documents publishable in a hypertext environment is a complex task requiring methods for segmentation, reorganization, and linking. The HyTex project, funded by the German Research Foundation (DFG), aims at the development of conversion strategies based on text-grammatical features. One focus of our work is on topic-based linking strategies using lexical chains, which can be regarded as partial text representations and form the basis of calculating topic views, an example of which is shown in Figure 1. This paper discusses the development of our lexical chainer, called GLexi, as well as several experiments on two aspects: Firstly, the manual annotation of lexical chains in German corpora of specialized text; secondly, the construction of topic views.

The principle of lexical chaining is based on the concept of lexical cohesion as described by Halliday and Hasan (1976). Morris and Hirst (1991) as well as Hirst and St-Onge (1998) developed a method of automatically calculating lexical chains by drawing on a thesaurus or word net. This method employs information on semantic relations between pairs of words as a connector, i.e. classical lexical semantic relations such as synonymy and hypernymy as well as complex combinations of these. Typically, the relations are calculated using a lexical semantic resource such as Princeton WordNet (e.g. Hirst and St-Onge (1998)), Roget's thesaurus (e.g. Morris and Hirst (1991)) or GermaNet (e.g. Mehler (2005) as well as Gurevych and Nahnsen (2005)). Hitherto, lexical chains have been successfully employed for 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)), topic detection/topic tracking (e.g. Carthy (2004)).

In order to formally evaluate the performance of a lexical chaining system in terms of precision and recall, a (preferably standardized and freely available) test set would be required. To our knowledge such a resource does not yet exist—neither for English nor for German. Therefore, we conducted several annotation experiments, which we intended to use for the evaluation of GLexi. These experiments are summarized in Section 2. The findings derived from our annotation experiments also led us to developing the highly modularized system architecture, shown in Figure 4, which provides interfaces in order to be able to integrate different pre-processing steps, semantic relatedness measures,

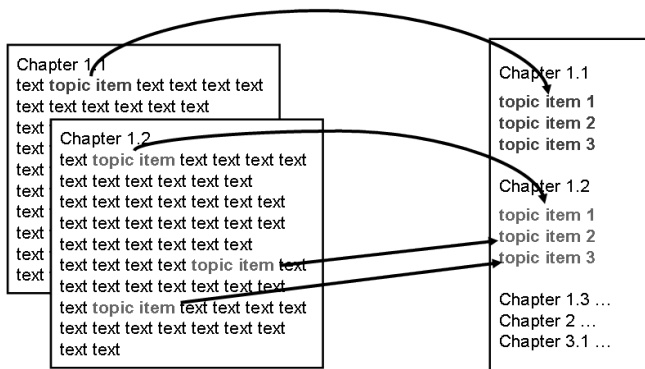


Abbildung 1: Construction of a Topic View Based on a Selection of Topic Items (= Thematically Central Lexical Unit) per Paragraph

resources and modules for the display of results. A survey of the architecture and the single modules is provided in Section 3. The challenges we experienced while annotating lexical chains brought us to analyze the performance of GLexi by means of a multi-level evaluation procedure, which is discussed in Section 4.

2 Annotation Experiments

The annotation experiments referred to below were originally intended to facilitate the development of annotation guidelines and thereby to promote the formulation of a gold standard for the evaluation of GLexi. However, the results of a preliminary study as well as the experiments detailed in the literature on English data (see, among others, Morris and Hirst (2005) as well as Beigman Klebanov (2005)) demonstrate that a satisfactory degree of inter-annotator agreement is not yet achieved in the manual annotation of lexical chains .

From our point of view, this is due to at least three aspects: Firstly, the subjects are focussed on building an individual understanding of the text, which obscures the various features that establish text cohesion, such as lexical cohesion or deixis. Secondly, the subjects also appear to struggle with differentiating between different features of textual cohesion. Particularly the anaphora and coreference resolution appears to be interacting strongly with lexical cohesion and thus the lexical chains in a text (see e.g. Stührenberg et al. (2007)). Thirdly, there is no consensus among researchers with respect to the semantic relations relevant in lexical chaining. It was therefore impossible to ensure a consistent annotation in regard to the relation types considered.

For this reason, all three experiments described in the following should be regarded as pilot studies. They were drafted and conducted with the aim of gaining more knowledge

on lexical chaining. We deemed as particularly important, which aspects of computing lexical chains or of their manual annotation respectively might be relevant for our application scenarios, namely, the construction of topic views. Contrastingly, it was of less importance, whether a satisfactory inter-annotator agreement could be achieved.

Therefore, the actual evaluation of GLexi was not conducted by means of the data, which were annotated in the experiments, but is rather based on an evaluation procedure that is detailed in Cramer and Finthammer (2008a) and sketched in Section 4. In altogether all three annotation experiments, we had subjects annotate lexical chains or pre-stages/parts of lexical chains within texts. The task of the three experiments may be summarized as follows:

- experiment 1: manual annotation of lexical chains;
- experiment 2: manual search for (direct and indirect) relations between words or synsets within GermaNet;
- experiment 3: manual annotation of lexical chains, represented as mind-maps.

In **experiment 1**, seven subjects (all subjects were second-year students of philology or linguistics with no background in lexicography and no knowledge in applied or computational linguistics) were asked to annotate lexical chains within three texts (two newspaper/ magazine articles, one from faz.net and unicum.de respectively, as well as a lexicon entry out of the German Wikipedia). For this purpose, the subjects were given a 15-minute oral introduction to the concepts of lexical cohesion and lexical chaining, including some notes on the theoretical background as described by Halliday and Hasan (1976) and some example chains. Subsequently, they had five minutes to ask clarification questions. The subjects were then given the following documents (partially depicted in Figure 2): a list of all nouns of the three texts, an evaluation questionnaire (for evaluating the relevance of the noun or phrase for their text comprehension), a template for generating the chains, a list of the relations to be considered, and a feedback questionnaire. Thereupon, the subjects were asked to complete the task as far as possible within one hour. In order to get an impression of the time necessary to annotate a certain amount of text we limited the amount of time.

Results - experiment 1: Nearly all subjects aborted the annotation before the set time exceeded. In fact, the subjects found the evaluation of the relevance of a noun to be comparatively easy, while they found the actual annotation of lexical chains to be rather difficult. Based on their divergent solution strategies in annotating lexical chains, the subjects may be subsumed into two groups (with three or four subjects each): the first group reinterpreted the task in so far, as they organized the nouns in nets (which they themselves called mind maps) rather than in chains. Subjects of the second group changed their strategies of chaining several times throughout the set time and in doing so crossed out previous versions in order to substitute them with improved ones (e.g. versions containing more or less entries or versions connected via other relations).

By means of a subsequent oral interview as well as the feedback questionnaire, we were able to identify the following main aspects of our subjects' criticism, which referred

i.e. the path between **Blume** (Engl. flower) and **Baum** (Engl. tree), which spans three steps in regard to GermaNet, namely hypernymy - hyponymy - hyponymy) by means of a graphic user interface for GermaNet (see Finthammer and Cramer (2008) for more information on this). We thus intended to account for the complaints by our subjects in experiment 1 that the semantic relation types did not suffice in order to satisfactorily complete the annotation of lexical chains. The subjects were given a fraction of the word pairs of experiment 1 and were asked to trace paths between these words with respect to GermaNet; they had a time-frame of four hours to complete the task.

Results - experiment 2: In principle, the following four constellations (see Cramer and Finthammer (2008b) for examples) could be identified:

- intuitively, a **semantic relation exists** between the words of the pair and this **connection can easily be identified** within GermaNet;
- intuitively, a **semantic relation exists** between the words of the pair, **but this relation can not easily or not at all be identified** within GermaNet;
- Intuitively, **no semantic relation exists** between the words of the pair, **but a short path** can easily be identified between the words within GermaNet;
- intuitively, **no semantic relation exists** between the words of the pair, **and no short path nor a path at all** can be identified within GermaNet.

In spite of the graphic user interface (see Finthammer and Cramer (2008)), there were almost no cases where the subjects were able to identify a path in GermaNet within an acceptable time-frame. In most cases, the search for a path terminated after two or three steps without any results. In these cases, the subjects were not able to decide intuitively on the next steps. Admittedly, it is not surprising that paths can only be detected manually with a great expenditure of time. But the results, that on the one hand even short paths run across inappropriate nodes (also see Cramer and Finthammer (2008b)) and that, on the other hand, intuitively, nodes being close to each other are only connected via long paths are markedly critical for the qualitative evaluation of word-net based relatedness measures.

In **experiment 3**, two subjects (of the above mentioned seven) were asked to construct lexical nets (similar to mind-maps) for three texts on the basis of the concept of lexical cohesion. We instructed them to consider the introduction they were given in experiment 1 as well as the results of the oral interviews and the feedback questionnaires. They first segmented the texts into paragraphs, for each of which one net was to be created. In a next step, the words and phrases of the paragraphs were transferred into net structures, which may be regarded as a partial representation of the textual content. An example of this can be found in Figure 3. As the figure illustrates, independently of each other, both subjects organized the words and phrases of the respective paragraph departing from a center (in regard to content).

The **results of the three annotation experiments** can be summarized as follows: The annotation of lexical chains within texts forms a complex task and is hardly viable

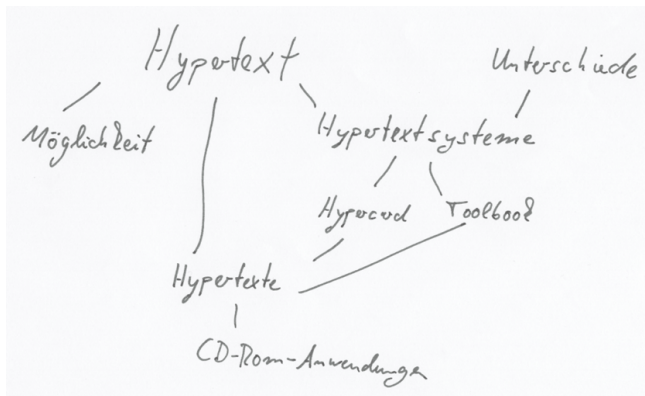


Abbildung 3: Example of a Manually Generated Net Structure as Complement or Substitute for Lexical Chains

along with achieving a sufficiently strong agreement of the subjects. These results correspond—in our opinion—to the results for English data as described in Beigman Klebanov (2005) as well as Morris and Hirst (2005). In a nutshell, developing sustainable annotation guidelines from the different experiments was ultimately impossible. Nevertheless, the results were relevant for our subsequent research on lexical chains for German data:

- Firstly, representing lexical chains as nets led us to the idea that the lexical units of a paragraph might be arranged around one or more words/word groups and thus around one or more thematic center(s) (we call them topic items). These topic items seem to feature a dense net structure and strong relations, which, in turn, forms the basis for the construction of topic views.
- Apart from this, the results emphasize that the performance of a system for calculating lexical chains cannot be evaluated by means of a manually annotated data. For this reason, an alternative approach for evaluation needed to be designed. Our suggestion for such an alternative procedure is sketched in Cramer (2008) and is briefly outlined in Section 4.

3 GLexi-Architecture

Drawing on the results of the previously described annotation experiments, we devised a modular system for calculating lexical chains/nets within German corpora. The basic modules of our system called GLexi (spoken: *galaxy*) are summarized in Figure 4. All modules are designed in such a way that the user of GLexi is able to additionally integrate own components, such as alternative pre-processing steps or resources. All

Tabelle 1: Options of Parameterization of GLexi Including a Compilation of the Configurations Used so far in the Experiments

Adjustable Parameters	Used Parameters
Pre-Processing: sentence boundary detection, tokenization, POS-tagging, morphological analysis, chunking	all pre-processing steps
Resources: GermaNet, GermaTermNet (see Beikwenger (2006) for more information on GermaTermNet), Google-API	GermaNet and GermaTermNet
Relatedness Measures: 8 based on GermaNet 3 based on Google (see e.g. Cilibrasi and Vitanyi (2007))	all 8 GermaNet based measures

options of parameterization—which were subject to our experiments up to now and which we use for calculating topic items (and topic views)—are compiled in Table 1.

The depiction of the system structure in Figure 4 also illustrates the chaining procedure: Based on the input (in an XML format particularly devised for this purpose) GLexi initially checks which units are to be considered as chaining candidates. Thereafter, all information on the candidates contained in the input is collected and hence is available for the core algorithm as well as the output generation. For each candidate pair GLexi then tests whether a suitable semantic relation can be identified on the basis of the selected resource and semantic relatedness measure. If this is the case, the pair is considered as a chaining pair and accordingly stored in the meta-chains¹ including its relatedness measure value. Having calculated the relatedness measure values for all possible pairs, i.e. having filled the meta-chains, the output of the results can be constructed. Again, different options are available: apart from the actual lexical chains (see e.g. the algorithm by Hirst and St-Onge (1998)) it is also possible to display all candidates including their relations as a net structure. An example of this is depicted in Figure 5. Obviously, we derived this format from the net structures as they were manually generated by our subjects in the annotation experiment 3 (see Section 2 and Figure 3). The net structure as a substitute for classical lexical chains, also forms the basis for calculating topic items (and topic views), as depicted in Figure 6.

4 GLexi–Evaluation

As mentioned above, no gold standard has been compiled so far for the evaluation of a lexical chainer and, in addition, the previously described results of the experiments illustrate that the manual annotation of such a gold standard represents yet unsolved challenges. We therefore suggest a four-step evaluation procedure as an alternative approach. A detailed discussion of this evaluation procedure is provided in Cramer and

¹ See Silber and McCoy (2002) for more information on the concept of meta-chains.

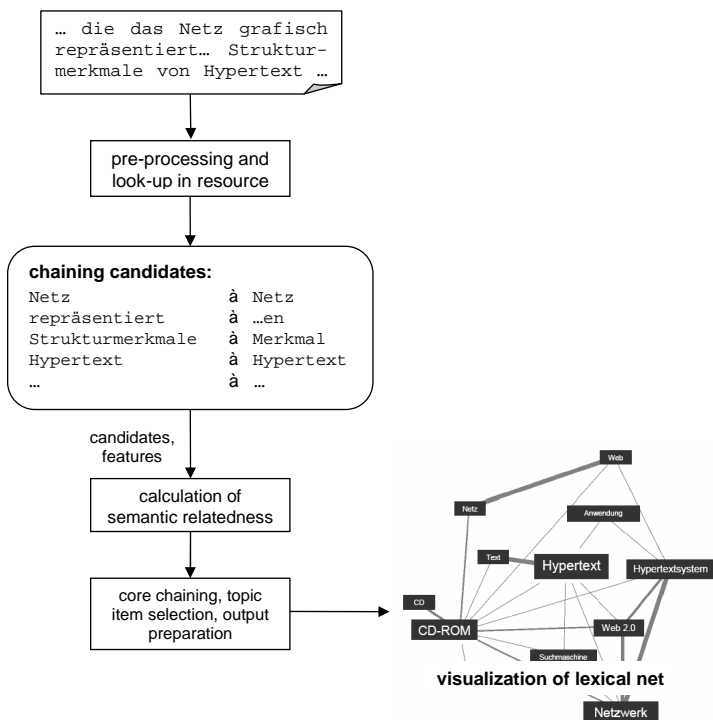


Abbildung 4: Architecture of GLexi

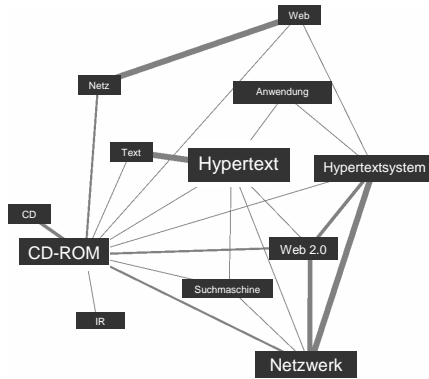


Abbildung 5: Example of a Lexical Net Generated by Means of GLexi

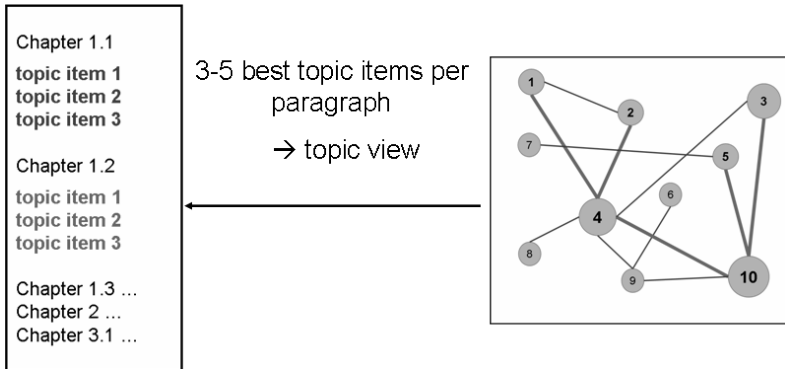


Abbildung 6: Output of GLexi as a Lexical Net Forms Basis for Calculating Topic Items and Topic Views: Choose the 3-5 Most Important Topic Items, Supplement TOC Accordingly.

Tabelle 2: Coverage of GermaNet in Regard to the HyTex Core Corpus

Approx. 29,000 (Noun) Tokens split into			
56% in GermaNet	44% not in GermaNet		
	15% inflected	12% compounds	17% proper names nominalization, abbreviation etc.

Finthammer (2008a). Therefore we limit the following description of the procedure to aspects relevant to the computation of topic views.

For evaluating GLexi, we drew on GermaNet (see e.g. Lemnitzer and Kunze (2002), version 5.0), the Google-API, and a word frequency list provided by S. Schulte im Walde² as resources for our eleven semantic relatedness measures. We additionally used parts of the HyTex core corpus (see Beißwenger and Wellinghoff (2006)), which we pre-processed by means of the Temis Tools³ and transformed into the previously mentioned XML format.

4.1 Coverage

With respect to the coverage of GLexi, two settings may be distinguished according to the resource used: If GermaNet forms the basis for calculating lexical chains, approximately 56% of the noun tokens in our corpus will be covered, see Table 2. If, in turn, the calculation of the semantic relatedness is based on the co-occurrence measures based on the Google-API, all words in the texts are accounted for. Having said that, using Google based relatedness measures involves two essential shortcomings: firstly, it does not provide a word sense disambiguation on-the-fly, as is the case using e.g. GermaNet; secondly, as the results given in Section 4.3 demonstrate, the correlation between the Google co-occurrence based measures and the average assessments of the subjects in regard to semantic relatedness still ranges below the measures which were achieved using the measures based on GermaNet.

4.2 Quality of Disambiguation

In order to evaluate the quality of word sense disambiguation, we manually annotated a fraction of the HyTex core corpus. As a next step, lexical chains were calculated for these data; in deciding upon the affiliation of a word (or a lexical unit of the corpus respectively) with a lexical chain, its word sense is simultaneously disambiguated. By comparing the decisions made on the basis of the chains calculated with the manual annotation, the quality of the disambiguation of GLexi may be assessed. Depending on

²We kindly thank Dr. Schulte im Walde for her support.

³For the experiments described here, the Insight DiscovererExtractor Version 2.1 was used. We also kindly thank the Temis group for supplying their software and for their support.

Tabelle 3: Correlation between Human Judgements and Relatedness Measure Values with Respect to the 100 Word Pairs

	Graph Path	Tree Path	Wu-Palmer	Leacock-Chodorow
correl.	0,41	0,42	0,36	0,48
	Hirst-StOnge	Resnik	Jiang-Conrath	Lin
correl.	0,47	0,44	0,45	0,48
	Google Quotient	Google NGD	Google PMI	
correl.	0,24	0,29	0,27	

the measure used, the results range between approximately 35% and 60%. In regard to the quality of their disambiguation, the measures introduced by Resnik (1995), Wu and Palmer (1994) and Lin (1998) perform best.

4.3 Quality of Semantic Relatedness Measures

In order to evaluate the performance of our eleven relatedness measures, we drew on a method typically employed in this context, namely, we compared the semantic relatedness measures values for a list of word pairs with human judgements of these pairs. Thus, the average assessments and the associated automatically calculated relatedness measure values for the word pairs are juxtaposed: Table 3 depicts the correlation between the human judgements and the eleven measures. Obviously, the measure values are scattered, which results in the rather low correlation coefficients. A detailed analysis of our human judgement experiments and a comparison with similar studies can be found in Cramer and Finthammer (2008a) and Cramer (2008).

4.4 Application Scenario

As mentioned above, the application scenario we aim at is the construction of topic views, an example of which is displayed in Figure 1. In order to automatically calculate topic views for given text passages, we mainly draw on the lexical nets generated by GLexi. We integrated the (in the following described) algorithm for the calculation of topic views as an additional module of GLexi: on the basis of the lexical nets we rank the words/phrases (topic item candidates) of a passage with respect to their topic strength; thus, we rank the candidates which are most relevant at the top of a topic item candidate list. The decision on the ranking of a given topic item candidate is mainly based on three feature types: firstly, the density of the lexical net for the given candidate, secondly, the strength of its relations, and, thirdly, its *tf/idf* score. We regard the top three to five (depending on the length of the passage) topic item candidates as the topic items of the given passage and construct the topic view by supplementing the topic items to the table of contents. In order to evaluate the performance of the above described algorithm, we drew on the manual annotation of topic items. Initial

annotation experiments show that an inter-annotator agreement of approximately 65% can be achieved. We also found that when evaluating the automatic calculation of topic views with respect to the manual annotated data, an overlap of 55% to 75% can be achieved. Our initial results also stress that GLexi is able to compute high quality topic views if the passages are of a certain length and if the topic item candidates are appropriately modeled in the lexical semantic resource employed. Interestingly, in spite of the moderate performance of GLexi with respect to its coverage, its word sense disambiguation performance and the semantic relatedness measures used, we were able to achieve—with only a few simple features—relatively good results in calculating topic views. However, we certainly need to systematically explore the calculation of topic views in a follow-up study.

5 Outlook

The results of our annotation experiment describe here as well as the evaluation of our system GLexi demonstrate that the concept of lexical chains as well as their automatic construction leaves a number of aspects unsettled: Firstly, it is questionable to what extent lexical chains may be distinguished from anaphoric structures or coreference respectively, or, put vice versa, how far these three concepts might be merged into a homogenic concept. Moreover, it remains unclear, whether we are dealing with lexical nets rather than lexical chains—as the subjects of experiment 1 stressed. The experiments on the construction of topic views however show that it might indeed be reasonable to replace the concept of lexical chains by a new concept of lexical nets. We therefore plan, as a follow-up study, to investigate the (basic) features of lexical nets and also intend to incorporate the findings of linguists on lexical (semantic) cohesion into this new concept more thoroughly. Secondly, the moderate performance of GLexi as detailed in Sections 4.1 to 4.3 indicates that lexical chaining (or lexical netting) might be a not yet well understood method for the construction of partial text representations. We find that particularly the quality of word sense disambiguation (which should—at least according to the theory of lexical chaining—be conducted on-the-fly while chaining words of a text) and the performance of the semantic relatedness measures do not meet our demands. The quality of disambiguation might well be improved by enhancing the pre-processing, but still the problem of calculating the semantic relatedness remains unsettled. The latter, again, consists of diverse sub-aspects: First of all, although there has been much research (see Morris and Hirst (2004) as well as Boyd-Graber et al. (2006)) on the question which types of semantic relations are actually relevant at all (for the calculation of lexical chains as well as in principle), we consider this issue unsettled. In addition, the human judgement experiments typically used in order to assess the performance of a semantic relatedness measure, are—in our opinion—not well understood, i.e. it is unclear what exactly is measured in such an experiment, and furthermore, the experimental set-up is not well defined. And finally, all measures which have been taken into account so far—do not consider those relations that arise exclusively from the content of the text and which can evolve within a text only. Despite these numerous unsettled questions, the

first application-based results demonstrate that lexical chains are convenient and helpful for the calculation of topic items and topic views. We therefore intend to systematically investigate—which parameter settings perform best for the calculation of topic views—and feel confident that we will—in the long run—be able to achieve results of high quality for our corpora of specialized text.

Acknowledgement

We kindly thank Angelika Storrer, Michael Beißwenger, Christiane Fellbaum, Claudia Kunze, Lothar Lemnitzer, Alexander Mehler, and Sabine Schulte im Walde for their support and valuable comments and suggestions. Moreover, we thank our dedicated subjects.

Literatur

- Barzilay, R. and Elhadad, M. (1997). Using lexical chains for text summarization. In *Proc. of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*.
- Beigman Klebanov, B. (2005). Using readers to identify lexical cohesive structures in texts. In *Proc. of ACL Student Research Workshop (ACL2005)*.
- Beißwenger, M. (2006). Termnet—ein terminologisches Wortnetz im Stile des Princeton Wordnet. Technical report, University of Dortmund, Germany.
- Beißwenger, M. and Wellinghoff, S. (2006). Inhalt und Zusammensetzung des Fachtextkorpus. Technical report, University of Dortmund, Germany.
- Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. (2006). Adding dense, weighted, connections to wordnet. In *Proceedings of the 3rd Global WordNet Meeting*, pages 29–35.
- Carthy, J. (2004). Lexical chains versus keywords for topic tracking. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science. Springer.
- Cilibrasi, R. and Vitanyi, P. M. B. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19.
- Cramer, I. (2008). How well do semantic relatedness measures perform? a meta-study. In *Proceedings of the Symposium on Semantics in Systems for Text Processing*.
- Cramer, I. and Finthammer, M. (2008a). An evaluation procedure for word net based lexical chaining: Methods and issues. In *Proceedings of the 4th Global WordNet Meeting*, pages 120–147.
- Cramer, I. and Finthammer, M. (2008b). Tools for exploring germanet in the context of cl-teaching. In *Text Resources and Lexical Knowledge. Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008*.
- Finthammer, M. and Cramer, I. (2008). Exploring and navigating: Tools for germanet. In *Proceedings of the 6th Language Resources and Evaluation Conference*.
- Green, S. J. (1999). Building hypertext links by computing semantic similarity. *IEEE Transactions on Knowledge and Data Engineering*, 11(5).

- Gurevych, I. and Nahnsen, T. (2005). Adapting lexical chaining to summarize conversational dialogues. In *Proc. of the Recent Advances in Natural Language Processing Conference (RANLP 2005)*.
- Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman, London.
- Hirst, G. and St-Onge, D. (1998). Lexical chains as representation of context for the detection and correction malapropisms. In Fellbaum, C., editor, *WordNet: An electronic lexical database*.
- Lemnitzer, L. and Kunze, C. (2002). Germanet - representation, visualization, application. In *Proc. of the Language Resources and Evaluation Conference (LREC2002)*.
- Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. of the 15th International Conference on Machine Learning*.
- Mehler, A. (2005). Lexical chaining as a source of text chaining. In *Proc. of the 1st Computational Systemic Functional Grammar Conference, Sydney*.
- Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1).
- Morris, J. and Hirst, G. (2004). Non-classical lexical semantic relations. In *Proc. of HLT-NAACL Workshop on Computational Lexical Semantics*.
- Morris, J. and Hirst, G. (2005). The subjectivity of lexical cohesion in text. In Chanahan, J. C., Qu, C., and Wiebe, J., editors, *Computing attitude and affect in text*. Springer.
- Novischi, A. and Moldovan, D. (2006). Question answering with lexical chains propagating verb arguments. In *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the IJCAI 1995*.
- Silber, G. H. and McCoy, K. F. (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4).
- Stührenberg, M., Goecke, D., Diewald, N., Mehler, A., and Cramer, I. (2007). Web-based annotation of anaphoric relations and lexical chains. In *Proc. of the Linguistic Annotation Workshop, ACL 2007*.
- Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*.