Using Data Mining and the CLARIN Infrastructure to Extend Corpus-based Linguistic Research

Thomas Bartz
TU Dortmund University
Department of German Language and Literature
44227 Dortmund, Germany
thomas.bartz@tu-dortmund.de

Christian Pölitz
TU Dortmund University
Artificial Intelligence Group
44227 Dortmund, Germany
christian.poelitz@tu-dortmund.de

Katharina Morik
TU Dortmund University
Artificial Intelligence Group
44227 Dortmund, Germany
katharina.morik@tu-dortmund.de

Angelika Storrer
Mannheim University
Department of German Philology
68131 Mannheim, Germany
astorrer@mail.uni-mannheim.de

Abstract

Large digital corpora of written language, such as those that are held by the CLARIN-D centers, provide excellent possibilities for linguistic research on authentic language data. Nonetheless, the large number of hits that can be retrieved from corpora often leads to challenges in concrete linguistic research settings. This is particularly the case, if the queried word-forms or constructions are (semantically) ambiguous. The joint project called 'Corpus-based Linguistic Research and Analysis Using Data Mining' ("Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining" – 'KobRA') is therefore underway to investigating the benefits and issues of using machine learning technologies in order to perform after-retrieval cleaning and disambiguation tasks automatically. The following article is an overview of the questions, methodologies and current results of the project, specifically in the scope of corpus-based lexicography/historical semantics. In this area, topic models were used in order to partition search result KWIC lists retrieved by querying various corpora for polysemous or homonym words by the individual meanings of these words.

1 Introduction and Project Background

Large digital corpora of written language, such as those that are held by the CLARIN-D centers, provide excellent possibilities for linguistic research on authentic language data (McEnery et al., 2006; Lüdeling and Kytö, 2008; Lüdeling and Kytö, 2009). The size of the corpora allows for remarkable insights into the distribution of notable language usage phenomena with respect to time and/or domain-specific aspects. Not the least thanks to the efforts being done in CLARIN, are analyzing and query tools becoming more and more sophisticated, and thus, enabling researchers to search for word forms or constructions and filter the results with regard to part of speech types or morphosyntactic aspects. Despite these advances, the large number of hits that can be retrieved from corpora often leads to challenges in concrete linguistic research settings. This is particularly the case, if the queried word forms or constructions are (semantically) ambiguous. Researchers in linguistics do not usually examine word forms, but instead the terms representing the relations of word forms and their meanings. It is for this reason that word form-based filtering carried out by the current query tools is insufficient in many cases and leads to an unpredictable number of false positives. Depending on the amount of data, in-

tense manual effort is required for cleaning and disambiguation tasks (Storrer, 2011). Many research questions cannot even be addressed for this reason.

The joint project called 'Corpus-based Linguistic Research and Analysis Using Data Mining' ("Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining" – 'KobRA') is therefore underway to investigating the benefits and issues of using machine learning technologies in order to perform after-retrieval cleaning and disambiguation tasks automatically. To this end, German linguists (Universities of Dortmund and Mannheim), computational linguists (Berlin-Brandenburg Academy of Sciences and Humanities – BBAW, Institute for the German Language – IDS, University of Tübingen) and computer scientists (University of Dortmund) closely collaborate on concrete corpus-based case studies in the fields of lexicography, diachronic linguistics and variational linguistics. The case studies reflect the research actually carried out in these fields and are related to the specific research activities of the project participants. Three major German corpus providers, all CLARIN-D centers (BBAW, IDS, Tübingen University; see above), take part in the project, providing the corpus data and also plan to integrate the project results into the existing infrastructures. The data mining processes, that are made available in RapidMiner (formerly: 'YALE', Mierswa et al., 2006), one of the most widely used data mining environments, operate on search result KWIC lists derived from the corpora. These go beyond a mere search and can be used in order to filter or structure the search results, as well as to simplify the further processing of the data, where necessary, to target specific research questions (e.g. through annotation).

The following article is an overview of the questions, methodologies and current results of the project, specifically in the scope of corpus-based lexicography/historical semantics. In this area, topic models were used in order to partition search result KWIC lists retrieved by querying various corpora for polysemous or homonym words by the individual meanings of these words. The utility and the conditions of these methods are illustrated based on case studies on example words that are of interest to a linguist. German homonyms and polysemes of various parts of speech are presented, that different corpora were queried for. As topic models operate independently from language, it was thought that this method would be suitable for languages other than German as well. Therefore, experiments were also run with English language data.

2 Scope: Research on the Semantic Change of Words

The semantic change of words is interesting for linguistics in two respects: lexicographers and historical semantic researchers. Lexicographers follow the evolution of words in order to construct adequate lexicographic descriptions for example in order to update existing dictionary entries (Storrer, 2011; Engelberg and Lemnitzer, 2009), while researchers in historical semantics explore the possibilities, conditions and consequences of semantic innovations (Fritz, 2012; Fritz 2005; Keller and Kirschbaum 2003). In both cases, the deciding factor in furthering knowledge is the availability of structured text-corpora that allow the use of a word to be tracked over broad time lines and genres. Although comprehensive synchronous and diachronic text corpora with meta data to occurrence dates and text types are available along with accessible retrieval and analysis tools, and especially in the framework of CLARIN, an extensive and automatic semantic annotation of corpora at current technological standards is not yet suitably possible (Storrer, 2011; Rayson and Stevenson, 2008). This means that until now corpus-based exploration of semantic changes of a word have to be manually disambiguated for individual detection. Therefore, the distribution and process of semantic change can presently only be described on the basis of few examples and a relatively small data corpus (Fritz 2005; Keller and Kirschbaum 2003).

3 Data Mining Approach: Disambiguation of KWIC Snippets

Instead of an exhaustive semantic annotation of large text corpora, it appears that it could be more promising to have a subsequent disambiguation of automatically generated KWIC snippets for a searched word retrieved via corpus query. This is also suggested through a series of preliminary results (see Section 4). Already a manual viewing of search results shows that the different meanings of a searched word are most easily recognized through the surrounding words. The usage of a word in a specific meaning evidently corresponds more frequently with occurrences of certain other words or linguistic structures in the environment of this word. Through data mining methods, this latent infor-

mation contained within a search result's context may be used for automatic disambiguation. For that purpose, all occurrences of a relevant word will be placed in context windows of a specific size and with help from word and co-occurrence statistics, distributions of the context words will be determined. These can then be regarded as representations of meanings. As a result, it will be possible to calculate the probability of the relevant word representing a certain meaning for every single context window. A major advantage of such methods that are inductively based on the related words' contexts is that this way unexpected or until now lexicographically unrecorded meanings can now be identified.

4 Related Work: Word Sense Induction (WSI)

The induction of semantic meaning in the area of data mining is already well researched. An early statistical approach was completed by Brown et al. (1991), Navigili (2009) provides a comprehensive overview on the current research. Brody and Lapata (2009) have shown that they obtained the best results with the help of Latent Dirichlet Allocation (LDA; Blei et al., 2003). In addition, they expanded their method to take into consideration various other context features besides the pure word occurrences (e.g. part of speech tags, syntax, etc.). Originally, LDA was used for thematic clustering of document collections. Navigli and Crisafulli (2010) have already shown this to also be useful for the disambiguation of small text snippets, for example when clustering the search results from a web search engine. Rohrdantz et al. (2011) showed the benefits of this method as a basis for the visualization of semantic change of example words from an English newspaper corpus, allowing them to observe the emergence of new meanings and reconstruct their development over time.

The approach proposed in this article differs from these previous works particularly through the application of LDA in search result KWIC snippets derived from queries in large text corpora. The results of a query in a (web) search engine usually correspond to (web) texts, which are closely connected thematically with the searched word. However, search results from a corpus search system are determined through the occurrences of the searched word throughout the corpus, regardless of the thematic relevance of the documents containing the words. In this way, the searched words generally occur in less normal and semantically less clear contexts. The text genre of belles-letters and of newspaper texts often include metaphorical usages. Based on Rohrdantz et al. (2011), KWIC snippets from different text type areas will be used, all of which – apart from one example – are in German.

The benefits and issues of using clustering methods like LDA for the automatization of disambiguation of the search results KWIC snippets derived from corpora are, as of yet, barely researched. In the context of CLARIN-D, there are corpora available to the KobRA project (details about queried corpora see Section 6), which include extensive linguistic (annotations of parts of speech and syntax) and document meta data (examples assigned to text genres and time frames). This is why the project also allows for insights relating to the questions of which attributes may improve the results of clustering methods, such as LDA, and how the KWIC snippets and attributes may ideally be represented for these methods.

5 Evaluated Data Mining Techniques and Environment

The data mining processes evaluated in the KobRA project are implemented as a plug-in in the data mining framework RapidMiner (formerly: 'YALE', Mierswa et al., 2006; see Figure 1). RapidMiner allows one to easily perform large scale data analysis and offers a plethora of methods to import, transform, and analyse textual data as well as to present and visualize the results of the analysis. Besides already available data mining methods for classification and clustering, additional methods were implemented for efficient feature extraction and calculation for large amounts of documents as well as for word sense disambiguation. The plug-in also includes methods to efficiently access linguistic data sources, as well as sophisticated methods to extract linguistic and document features (if available) from KWIC lists. An integrated annotation environment enables linguists to add additional annotations to the KWIC snippets and the words retrieved from the data sources.

For the disambiguation approach described and evaluated in this paper (see Sections 3, 7, 8), we implemented the Latent Dirichlet Allocation method (LDA; Blei et al., 2003; Steyvers et al., 2004; Blei and Lafferty, 2006). LDA models the probability distributions of the words and the snippets from the corpus query result lists. The probability distributions are scattered over a number of what are known as latent topics that correspond to different meanings of a queried word. Based on the words and word

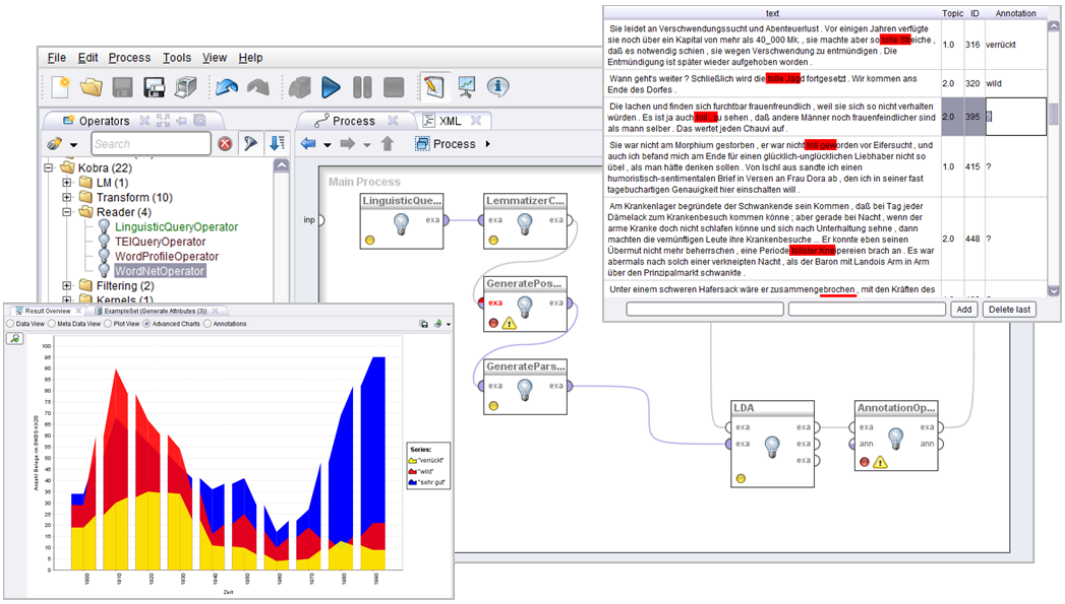

Figure 1: RapidMiner data mining framework and project plug-in.

co-occurrences in the snippets, LDA assigns those words that appear together into the same topics. These topics are then interpreted as meanings.

The probability distributions of the topics for a given word or snippet are multinomial distributions φ respectively θ. These distributions are drawn from a Dirichlet distribution Dir(β) respectively Dir(α) for the meta parameter α and β. The Dirichlet distribution is a distribution over distributions.

The estimation of the distributions is done via a Gibbs sampler as proposed by Griffiths and Steyvers (2004). The Gibbs sampler models the process of assigning a word or snippet to a certain topic based on the topic distributions of all other topics. This is a Markov chain process and converges to the true topic distributions for given words and snippets.

An important aspect, that is investigated, is the possibility to integrate further information into the generation of the topic models. Steyvers et al. (2004), for instance, integrate additional information like authorships of documents in the topic models. We use their approach to integrate information about the text genre classes the query result snippets are attributed to. This enables an additional investigation of how topics, words and snippets distribute over these classes. Moreover, the integration of the publication dates provided with the snippets are of interest to this study. Blei and Lafferty (2006), for example, introduced a dynamic topic model that facilitates analyzing the development of the found topics over time.

6 Words of Interest and Queried Corpora

For the case studies outlined in this article, we queried various corpora for a choice of words that are linguistically interesting, because they recently or over a long period of time have taken on new meanings, or their original meanings have changed. According to the assumed time period of the meaning changes, different corpora were queried. Moreover, we chose example words belonging to different parts of speech. Using this setting, we expect interesting insights in possible corpus- or word class-specific differences in the usefulness of the evaluated data mining techniques. The below examples are the basis for the following outlined experiments. Details about the corpora used can be found subsequently.

- Through the technical innovation of the 20th century, the noun “Platte” had a pronounced differentiation in its range of meanings. Along with the meaning *flaches Werstück* (flat workpiece) or *Teller* (plate), different uses gradually appeared: *fotografische Platte* (photographic plate), *Schallplatte/CD* (gramophone record/compact disk) oder *Festplatte* (hard disk). A search for the lemma of “Platte” in the DWDS core corpus of the 20th century results in 2886 KWIC snippets.

- During the commercial distrubition of the telephone in the 20s and 30s of the 20th century, a new meaning appeared for the verb “anrufen“ besides its original meaning *rufen/bitten* (to cry/to appeal to someone): that of *telefonieren* (to telephone). A search for the verb “anrufen“ in the DWDS core corpus of the 20th century results in 2085 KWIC snippets.
- Since the financial and bank crisis (circa 2007) the noun “Heuschrecke” has a new use, along with its original meaning *Grashüpfer* (locust), to now also describe persons involved in what is known as “Heuschreckenkapitalismus” (locusts capitalism). A search for “Heuschrecke“ in the DWDS newspaper corpus “Die ZEIT” results in 715 KWIC snippets.
- The adjective “zeitnah“ appears to have received a new prototypical meaning in the last 20 to 30 years, *unverzüglich* (prompt), while still retaining its original meaning of *zeitgenössisch* (contemporary)/*zeitkritisch* (critical of the times). A search for “zeitnah“ in the DWDS newspaper corpus “Die ZEIT” results in 597 KWIC snippets.
- The adjective “toll“ has had a remarkable meaning shift in the last century; its original meaning of *irre* (insane) changed to *ausgelassen/wild* (jolly/wild) and to its now positively attributed meaning *sehr gut* (very good). A search for the adjective “toll“ in the Tübingen Treebank of Diachronic German results in 5793 KWIC snippets, and a corresponding search in the DWDS core corpus of the 20th century results in 1745 KWIC snippets.
- The conjunction “da“ (as/because) was almost only used for temporal meaning in early records. Today, it is mostly used causally. A search for the conjunction “da“ in the Tübingen Treebank of Diachronic German results in 123496 KWIC snippets.
- The choice of the English noun “cloud” represents this aricle’s first attempt to use the proposed approach on none German language data. A new meaning appears to have evolved in the last decades with the emergence of large computer networks (clouds), next to the original meaning (*mass of condensed water, smoke, dust or other elements*). A search for “cloud“ in the corpora of the Leipzig Corpora Collection results in 1486 KWIC snippets.

The DWDS core corpus of the 20th century (DWDS-KK), constructed at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), contains approximately 100 million running words, balanced chronologically (over the decades of the 20th century) and by text genre (belles-lettres, newspaper, scientific and functional texts). The newspaper corpus “Die ZEIT” (ZEIT) covers all the issues of the German weekly newspaper “Die ZEIT” from 1946 to 2009, approximately 460 million running words (Klein and Geyken, 2010; Geyken, 2007).

The Tübingen Treebank of Diachronic German (TüBa-D/DC) is a syntactically annotated (constituent parse trees) corpus of selected diachronic language data from the German Gutenberg Project (http://gutenberg.spiegel.de/), a community-driven initiative of volunteers making copyright-free belles-lettres from 1210 to 1930 publicly available via web-interface. TüBa-D/DC that is hosted by the CLARIN-D Center at the Eberhard Karls University of Tübingen contains approximately 250 million running words (Hinrichs and Zastrow, 2012).

The Leipzig Corpora Collection (LCC) consists of corpora in different languages that contain randomly selected sentences from newspaper texts and a web sample (Quasthoff et al., 2006). We used the English corpus with language data from newspapers and the English Wikipedia, covering the time span from 2005 to 2010.

The corpus queries provide KWIC text snippets with occurrences of the investigated words (with including inflected forms). In addition, the publication dates and other document metadata (for the TüBa-D/DC: titles, author names; for the DWDS-KK: text genre classes) are given for each snippet.

7 Experiments and Evaluation

Accounting for our research question for the optimal representation of the KWIC snippets and our selection of example words and corpora (Section 6), eight evaluative treatments of the approach outlined in Section 5 were created. These can be systematically separated into the following aspects:

- **Queried word and part of speech:** noun, verb, adjective, or conjunction

- **Number of meanings:** two or more
- **Queried corpus:** corpus of contemporary German (DWDS-KK, ZEIT) or diachronic corpus (TüBa-D/DC, orthographically normalized)
- **Language of the corpus:** German or English
- **Number of KWIC-Snippets:** More or less than 1000 snippets

In addition, every treatment was tested to check which context size (20, 30, or 40 words) led to the best results for the relevant word. The following Table 1 shows an overview of the evaluative treatments for the outlined data mining techniques in Section 5.

Treatment	Word	Part of Speech	Meanings	Corpus	Language	Snippets	Context		
							20	30	40
1	Platte	noun	5	contemporary	German	> 1000	X	X	X
2	toll	adjective	3	contemporary	German	> 1000	X	X	X
3	anrufen	verb	2	contemporary	German	> 1000	X	X	X
4	Heuschrecke	noun	2	contemporary	German	< 1000	X	X	X
5	zeitnah	adjective	2	contemporary	German	< 1000	X	X	X
6	toll	adjective	2	diachronic	German	> 1000	X	X	X
7	da	conjunction	2	diachronic	German	> 1000	X	X	X
8	cloud	noun	3	contemporary	English	> 1000	X	X	X

Table 1: Evaluation treatments.

For the evaluation purposes, 30 percent of the retrieved KWIC snippets for the queried words were disambiguated manually by two independent annotators. Table 2 shows the obtained inter-annotator-agreement (kappa: Cohen, 1960):

Treatment	Word	Agreement
1	Platte	0.82
2	toll	0.76
3	anrufen	0.97
4	Heuschrecke	0.98
5	zeitnah	0.91
6	toll	0.71
7	da	0.75
8	cloud	0.92

Table 2: Inter-annotator-agreement of the manual disambiguation.

The automatic disambiguation approach was evaluated based on the manually disambiguated data sets. Therefore, topic models (see Section 5) were generated to extract the meanings of the queried words' occurrences and the results were compared to the labels attributed by the annotators. As a measure of reliability for the automatic disambiguation, we use one of the standard measures used to estimate the goodness of a word-sense disambiguation result, the F_1 score. The F_1 score is the weighted average of the disambiguation results' precision and recall in relation to the given annotations. This and further evaluation methods are described by Navigli and Vanella (2013).

8 Evaluation Results

8.1 Reliability of the automatic disambiguation using LDA

The following tables show the results achieved using the above described approach. The Tables 3-8 list the evaluation scores for the investigated treatments:

"Platte"		flat workpiece	plate	photographic plate	gramophone record/compact disk	hard disk
F_1 for context (words)	20	0.800	0.800	0.667	0.287	0.857
	30	0.998	0.875	0.500	0.381	0.988
	40	0.733	0.600	0.750	0.353	0.800

Table 3: Results for treatment 1.

"toll"		insane	jolly/wild	very good
F_1 for context (words)	20	0.519	0.571	0.167
	30	0.714	0.615	0.632
	40	0.625	0.667	0.500

Table 4: Results for treatment 2.

"anrufen"		to cry/to appeal to someone	to telephone
F_1 for context (words)	20	0.727	0.667
	30	0.800	0.800
	40	0.909	0.889

Table 5: Results for treatment 3.

"Heuschrecke"		locust	person
F_1 for context (words)	20	0.857	0.842
	30	0.800	0.933
	40	0.667	0.727

Table 6: Results for treatment 4.

"zeitnah"		prompt	contemporary/critical of the times
F_1 for context (words)	20	0.727	0.667
	30	0.888	0.800
	40	0.895	0.818

Table 7: Results for treatment 5.

"toll"		insane	jolly/wild
F_1 for context (words)	20	0.526	0.571
	30	0.625	0.750
	40	0.556	0.636

Table 8: Results for treatment 6.

"da"		temporal	causally
F_1 for context (words)	20	0.471	0.556
	30	0.353	0.529
	40	0.400	0.611

Table 9: Results for treatment 7.

"cloud"		mass of condensed water, etc.	computer network	name
F_1 for context (words)	20	0.526	0.500	0.471
	30	0.783	0.631	0.615
	40	0.467	0.545	0.684

Table 10: Results for treatment 8.

The results demonstrate that the advised task of automatic disambiguation of KWIC snippets retrieved from corpus queries (see Section 3) yield highly positive outcomes using the approach outlined above (see Section 5). In the best case scenario the average F_1 scores for the reliability of the method is around 0.732. However, depending on the searched word and preferred meaning the values varied in the range between 0,381 and 0,998 (again in the best case scenario). The generality of this method is therefore difficult to hypothesize. Still, according to the above formulated systematization of differences in the treatments (see Section 7) the following trends were established:

Word Form

It appears that the automatic disambiguation of nouns, verbs, and adjectives of the examined examples had, essentially, the same success rates. Similarly good values resulted from the example "Heuschrecke" (see Table 6) as with "zeitnah" (see Table 7) or "anrufen" (see Table 5). Nouns had the highest values (see also Table 3). The finer meaning differences of the conjunction "da" were not satisfactorily recognizable (see Table 9). The method is most promising in terms of content words. This is to be expected because of their function as semantic references. The applicability of this method in relation to grammatical words should be further investigated.

Number of Meanings

It appears, however, that the number of meanings of the examined examples systematically influenced the results. The method revealed lower results for "toll" (see Table 4) and "cloud" (see Table 10) than for examples that had only two meanings. This was also true for single meanings of "Platte" (see Table 3), while for the others the highest values were obtained. In essence, it appears that various meanings are differently identifiable.

Corpus and Language

At first glance, it appears that the chosen corpora (contemporary German vs. diachronic, German vs. English) had relatively similar results with the automatic disambiguation. The snippets' results for "toll" from the DWDS-KK (see Table 4) are comparable to those from TüBa-D/DC (see Table 8); this was also true for the English example "cloud" (see Table 10). In this respect, the evaluative success was expected as the texts from the TüBa-D/DC all lie within the orthographic normative form. More study is needed to determine if this method is also suitable for diachronic corpora with non-normative orthographic language data.

Number of Snippets and Size of the Contexts

Although the number of KWIC snippets used (500-1000 vs. 1000-5000) for each example appeared to have no systematic effect on the results – "zeitnah" (see Table 7) and "Heuschrecke" (see Table 6) were similarly well disambiguated, as were "Platte" (see Table 3), "toll" (see Table 8) or "anrufen" (see Table 5) – it was demonstrated that the method was most useful when the range of the contexts was 30 words before and after the examined word. Yet, it appears that for the verb "anrufen" (see Table 5) the most promising results came from the largest context. A reason for this could be that the verb in its function is more correlated with the sentence as a larger unit, while nouns and adjectives are already specified by their proximal contexts. This is supported by the slightly better results from the primarily adverbial used "zeitnah" (see Table 7) in the treatment with a context of 40 words.

8.2 Application for Research on the Semantic Change of Words

Using the automatic disambiguation easily allows the occurrences of single meanings of examined words to be identified and visualized. From the figures (see Figure 2-6) one can see the benefits of the integration of the query snippet's publication dates into the generation of the topic models: Researchers investigating semantic change are enabled to easily track the use of disambiguated word forms over time:

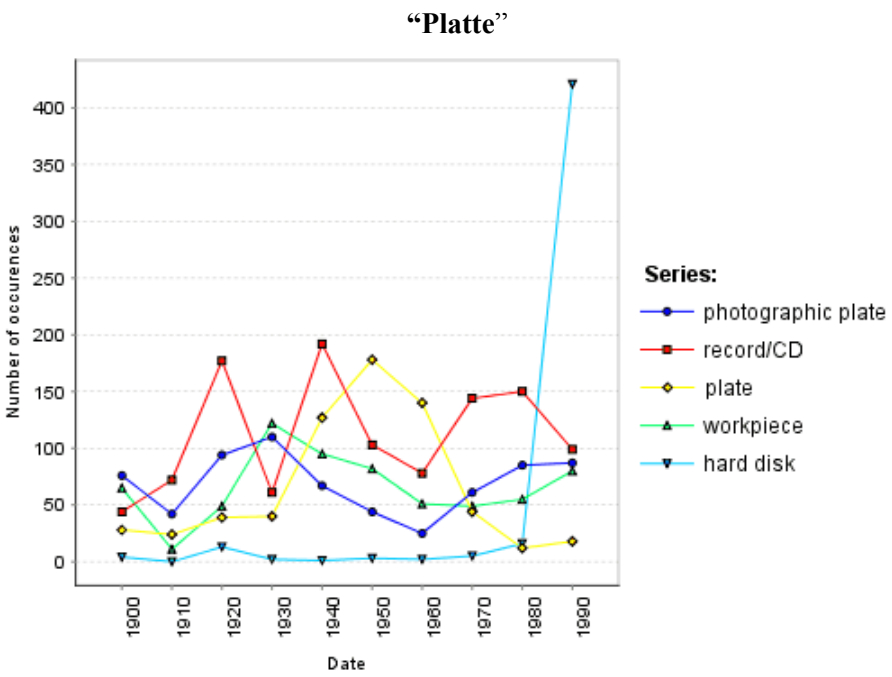

Figure 2: Occurrences of the word "Platte" with the meanings *flat workpiece*, *plate*, *photographic plate*, *gramophone record/compact disk*, *hard disk* in the decades of the 20th century.

The evolution of the meanings of "Platte" is illustrated traceable by Figure 2. The use of the meaning *hard disk* increased dramatically in the 90s, while the other meanings had a more uniform increase in usage in the different phases. The phases of more prevalent usage (e.g. the meaning *plate* in the 40s-60s or the meaning *photographic plate* in the 80s and 90s) are grounds for a more exact studies that would take the underlying KWIC snippets into account. With this in mind, the development of interactive visualisation, which is linked with the corpus base, would further simplify corpus based research on semantic change.

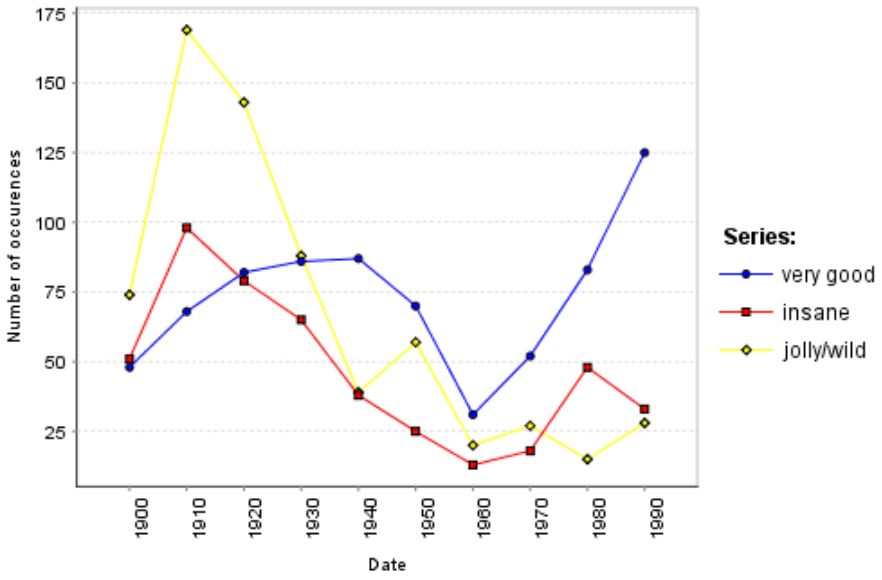

Figure 3: Occurrences of the word "toll" with the meanings *insane, jolly/wild, very good* in the decades of the 20th century.

Figure 3 clearly displays the semantic development of the word “toll“ during the 20th century. To the same degree that the older meanings *insane* and *jolly/wild* dropped in frequency, so did the newer meaning *very good* become more and more prominent.

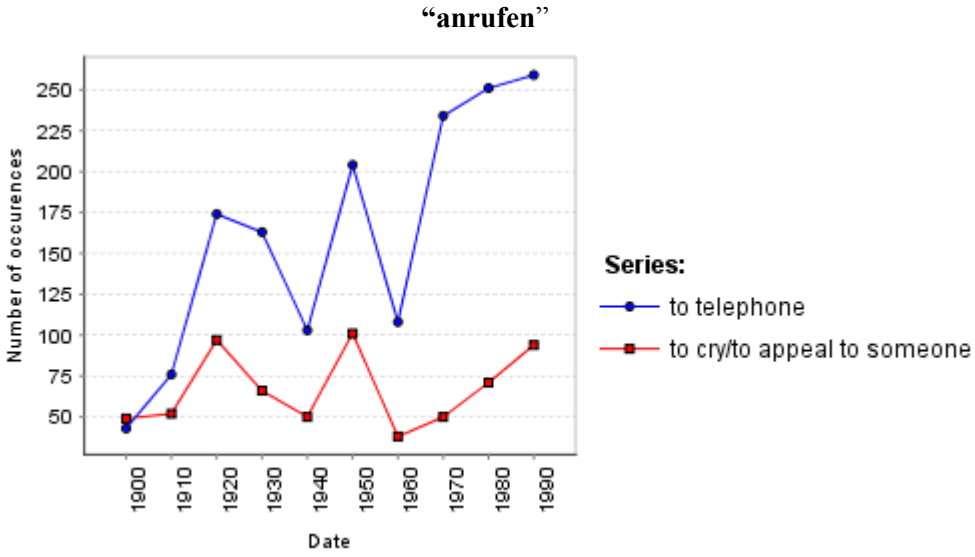

Figure 4: Occurrences of the word “anrufen” with the meanings *to cry/to appeal to someone, to telephone* in the decades of the 20th century.

Figure 4 shows that the strong increase in the use of the word “anrufen“ with the meaning *to telephone* occurred parallel to the commercial spread of telephones. The serrated frequencies that appear for both meanings between 1930 and 1970 could point to an irregularity in the balance of the corpus basis. This, again, underscores the need for a closer investigation of the underlying KWIC snippets.

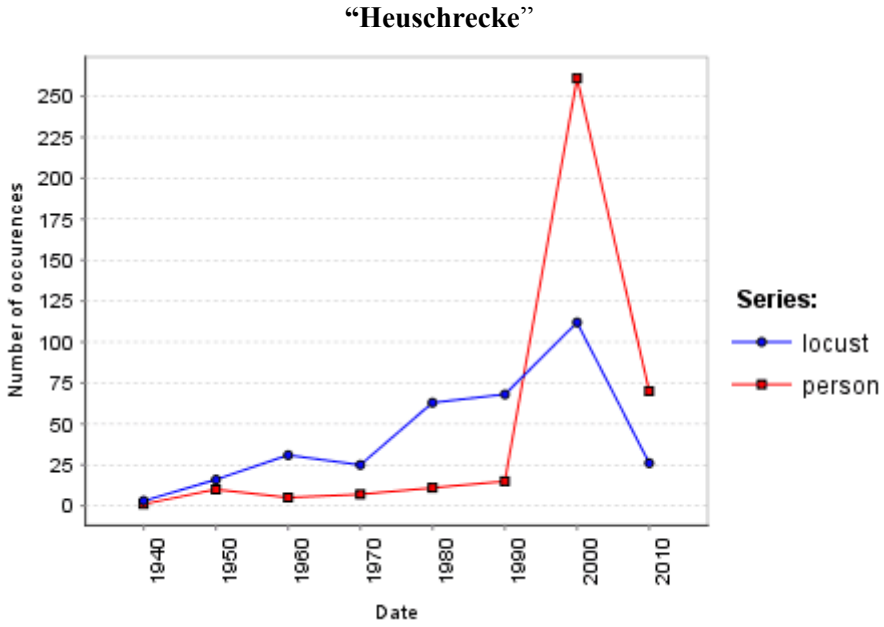

Figure 5: Occurrences of the word “Heuschrecke” with the meanings *locust, person* in the time span 1940-2010.

Figure 5 clearly shows a dramatic increase in the use of “Heuschreke“ with the meaning *person* in the 2000s, in the decades during the international financial and bank crisis. In the decade of the 2010s, there is a noticeable decline in the frequency of this use. However, the markedly smaller amount of records for this decade, in contrast to the others, could explain this discrepancy.

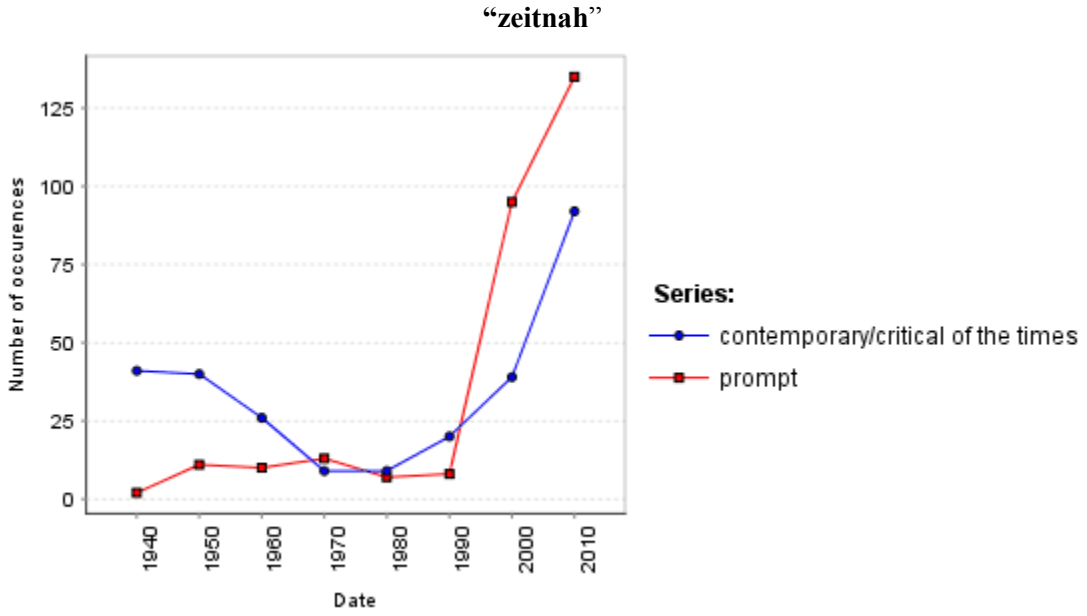

Figure 6: Occurrences of the word "zeitnah" with the meanings *prompt, contemporary/critical of the times* in the time span 1940-2010.

Finally, Figure 6 shows the sudden development, starting in the 2000s, of the meaning *prompt* as a new prototypical meaning for "zeitnah". What is interesting about this, is that at the same time there is a rise in the use of its older meaning *contemporary/critical of the times*. In order to check if this is accurate, or if this is just a cumulation of false meaning associations, it would again be advantageous to have the possibility of directly and interactively accessing the KWIC snippets.

9 Conclusion

The preceding report is a summary of questions, methods and selected results of the joint project called 'Corpus-based Linguistic Research and Analysis Using Data Mining' ("Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining" – 'KobRA') where German linguists, computational linguists and computer scientists closely cooperate in order to investigate benefits and issues of using data mining techniques for the automation of after-retrieval cleaning and disambiguation tasks in the area of corpus-based empirical language research. The methods used and evaluated in this project will be available for research and teaching within the data mining environment RapidMiner and from the CLARIN-D infrastructure.

This article was based mostly on the requirements and issues in the area of corpus-based lexicography/historical semantics. In this area, topic models were used in order to partition KWIC lists retrieved by querying various corpora for a choice of polysemous or homonym words according to the single meanings of the searched words. The reliability of the automatic method was evaluated with help from two independent annotators who manually disambiguated the evaluation data sets.

Overall, the evaluation gave positive results. The automatic disambiguation performed with similar success for content words such as nouns, verbs or adjectives. It is still to be seen if the usefulness of this method can be extended to grammatical words; for that, more study is needed. The number of meanings of each search word was found to impact the values of the results (less definitions, better results). In most cases it also appeared true that a medium sized context for the relevant word led to the best results. Neither the number of considered KWIC snippets in the range of 500-5000 nor the use of different (orthographically normalized) corpora had any noticeable effect on the results of the automatic disambiguation. More studies are needed to review the performance of this method for diachronic corpora with non-normative orthography.

Using the automatic disambiguation easily allows the occurrences of single definitions of examined words to be identified and visualized. The integration of the query snippet's publication dates enables researchers investigating semantic change to easily track the use of disambiguated word forms over

time. A next step of innovation for this method would be the development and testing of interactive visualizations, which would allow for direct access to the underlying corpus basis.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research ("Bundesministerium für Bildung und Forschung") in the funding line for the eHumanities [01UG1245A-E].

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3* (3), 993-1022.

David M. Blei and John D. Lafferty. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, 113-120.

Samuel Brody and Mirella Lapata. (2009). Bayesian word sense induction. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 103-111.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, 264–270.

Jacob Cohen. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement* 20, 37-46.

Stefan Engelberg and Lothar Lemnitzer. (2009). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Tony McEnery, Richard Xiao, and Yukio Tono. (2006). *Corpus-Based Language Studies – an advanced resource book.* London: Routledge.

Gerd Fritz. (2012). Theories of meaning change – an overview. In C. Maienborn et al. (Eds.), *Semantics. An International Handbook of Natural Language Meaning*. Volume 3. Berlin: de Gruyter, 2625-2651.

Gerd Fritz. (2005). *Einführung in die historische Semantik*. Tübingen: Niemeyer.

Alexander Geyken. (2007). The DWDS corpus. A reference corpus for the German language of the twentieth century. In C. Fellbaum (Ed.), *Idioms and collocations. Corpus-based linguistic and lexicographic studies*. London: Continuum, 23-40.

Thomas L. Griffiths and Mark Steyvers. (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences, 101* (Suppl. 1), 5228-5235.

Erhard Hinrichs and Thomas Zastrow. (2012). Automatic Annotation and Manual Evaluation of the Diachronic German Corpus TüBa-D/DC. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 1622-1627.

Rudi Keller and Ilja Kirschbaum. (2003). *Bedeutungswandel. Eine Einführung*. Berlin: de Gruyter.

Dan Klein & Christopher D. Manning (2003): Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1, ACL '03, pag-es 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wolfgang Klein and Alexander Geyken. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In U. Heid et al. (Eds.), *Lexikographica.* Berlin: de Gruyter, 79-93.

Anke Lüdeling and Merja Kytö. (Eds.). (2008). *Corpus Linguistics. An International Handbook.* Volume 1. Berlin: de Gruyter.

Anke Lüdeling and Merja Kytö. (Eds.). (2009). *Corpus Linguistics. An International Handbook.* Volume 2. Berlin: de Gruyter.

Ingo Mierswa et al. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*.

Roberto Navigli. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys, 41* (2), 10:1-10:69.

Roberto Navigli and Giuseppe Crisafulli. (2010). Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 116-126.

Roberto Navigli and Daniele Vannella. (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*, 193-201.

Uwe Quasthoff, Matthias Richter, and Chris Biemann. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, 1799-1802.

Christian Rohrdantz et al. (2011). Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 305-310.

Paul Rayson and Mark Stevenson. (2008). Sense and semantic tagging. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics.* Volume 1. Berlin: de Gruyter, 564-578.

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining*, 306–315.

Angelika Storrer. (2011). Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In K. Knapp et al. (Eds.), *Angewandte Linguistik. Ein Lehrbuch.* 3. vollst. überarb. und erw. Aufl. Tübingen: Francke, 216-239.