Erschienen in: Journal for Language Technology and Computational Linguistics Vol. 26 (2011) Nr. 2, S. 117–128.

Cerstin Mahlow, Britta Juska-Bacher

Exploring New High German texts for evidence of phrasemes

Most dictionaries containing phraseological information are restricted to a synchronic perspective. Diachronic information on structural, semantic, and pragmatic change over time has to be reconstructed by a time-consuming consultation of various dictionaries providing only punctual insights. In the OLdPhras, project we construct an online dictionary for diachronic phraseology in German from ca. 1650 to the present by combining dictionary exploration with corpus-based methods. This paper highlights some challenges we have met: How to select the "interesting" phrasemes, i.e., those that underwent some change? How to deal with historical corpora? How to include different kinds of phraseme variation? We present a semi-automatic corpus-based approach for the investigation of phraseme development. We argue for a combination of dictionary exploration and corpus-based methods to provide reliable and extensive information about the diachronic development of German phrasemes.

1 Introduction

Phraseology as a subfield of linguistics investigates form, meaning, use, and change of phrasemes (also referred to as phraseological units, idioms, or set phrases). Phrasemes are defined by polylexicality, relative stability, and idiomaticity (Burger, 2010, 36ff). Dictionaries—whether printed or electronic ones—usually describe the characteristics of phrasemes at a certain point in time, i.e., they are restricted to a synchronic perspective. For example, in a contemporary dictionary of German phraseology, one finds information on the current meaning of phrasemes such as *gegen den Strom schwimmen* ("to swim against the current", i.e., 'to oppose the opinion or the habits of the majority') and an example. General-purpose dictionaries give fewer explanations on a phraseme's development; etymological dictionaries like Kluge (2002) only provide information on the development of single words, not of multi-word units.

Metalexicographers have repeatedly criticized the neglect or the unsystematic presentation and placement of phrasemes in general and phraseological dictionaries (Kühn, 2003; Stantcheva, 2003; Burger, 2010). Often users cannot determine whether an example given represents established usage and can be found in real-world texts or whether it was made up by the author of the dictionary. In current phraseography research, empirical methods, i.e., analyzing large corpora, are emphasized to overcome some of these problem (see for example Mellado Blanco, 2009).

The project "German Proverbs and idioms in language change. Online-dictionary for diachronic phraseology (OLdPhras)" (started August 2010) funded by the Swiss National Science Foundation aims to provide information on the development—i.e.,

structural, semantic, and pragmatic change—of German phrasemes from ca. 1650 to the present. The resulting electronic dictionary is intended to serve researchers as a resource for further investigations as well as interested laypersons for information purposes. Alongside common lexicographic information on lexical units and their grammatical structure, we focus on evidence concerning lexical and semantic variants of phrasemes, and on changes in meaning and use by offering authentic examples from (New) High German texts. Comments on diachronic development will be based primarily on evidence from corpora considering observable characteristics with respect to lexical units, syntactic and morphosyntactic properties, semantic concepts, or pragmatic aspects. Additionally, we consider synchronic information from existing dictionaries covering that period and we describe usage and meaning at the time of a specific evidence, see Juska-Bacher and Mahlow (2012) for an example of results to be expected by using the example of *gegen den Strom schwimmen*.

In this paper we first look at the state of the art with respect to handling German phrasemes and refer to related work. Then we comment on the resources used in the OLdPhras project and outline specific challenges. We present our approach, with particular attention on how to overcome some of the obstacles due to the diachronic perspective, and report on first findings; as this is still ongoing work, we cannot provide extensive evaluations.

2 State of the art and related work

Defining phrasemes as *non-Fregian collocations*, we can search texts for potential collocations, which then have to be classified by phraseologists with respect to idiomaticity, resulting in a semi-automatic process. In our project we face two of the challenges pointed out by Rothkegel (2007, 1027): exploring which phrasemes are used in which forms and variants, and automatically identifying phrasemes in texts.

Fritzinger et al. (2009) propose the extraction of potential collocations for German to be based on fully syntactically parsed text. However, the implementation seems to be prototypical only. Seretan and Wehrli (2010) report on the extraction of collocations as preprocessing step to make the work of lexicographers easier. Rothkegel (2007) and Heid (2007) present various attempts to extract collocations from texts in different languages, Evert (2005) compares various approaches. A common feature is the attempt to reach high recall—to not miss a potential phraseme—resulting in rather low precision, requiring manual efforts to identify phrasemes.

All studies try to identify whether there are *any* collocations in a given text; there is no previous assumption on what to expect. Applying these approaches would be a completely corpus-driven method. In contrast, in the OLdPhras project, we follow a rather corpus-based approach—given a set of "interesting" phrasemes (see section 5.1) we explore corpora for evidence.

However, the approaches and methods developed so far are applied to *modern* texts. Annotated databases like Kuiper et al. (2003) describe phrasemes at a certain point in time, which then can be used in NLP tools. For our project with a strong diachronic

focus, two major issues arise: First, methods and resources developed for or trained on modern texts cannot be applied to texts from older language stages—there are differences with respect to orthography, lexicon, morphology, as well as syntax. Second, existing electronic resources like dictionaries or lexical databases (e.g., Kuiper et al. (2003)) reflect the *current* state of a language. We can distinguish compositional and non-compositional phrasemes in today's language use, but our interest is the point in time when arbitrary multi-word units started to be used in an idiomatic way rather than literally, or when certain kinds of variation of a phraseme were not used or not even allowed anymore. Synchronic resources (e.g., collections, tools) may thus only serve as a starting point. The lexicon developed by Keil (1997) for NLP purposes and used in experiments by Fischer and Keil (1996) seems to be no longer available; however, the proposed structure is of interest for our purposes.

The electronic resource most closely related to the aims of our project is the *Idiomdatenbank*, developed between 2003 and 2006 at the Berlin-Brandenburgische Akademie der Wissenschaften, see Fellbaum (2007); Fellbaum and Geyken (2005). However, *OLdPhras* is not simply an extension, but poses specific challenges related to a very simple fact: The time period to be investigated is longer (350 years vs. 100 years). During 350 years, more variation and change can be expected than during 100 years—more variation is assumed the further back in time we go (Burger and Linke, 1998). Change might occur on all levels relevant to a phraseme: (a) spelling, (b) lexical components, (c) syntactical structure of the multi-word unit, (d) semantics of single units, (e) semantics of the multi-word unit, (f) pragmatics. Taking as many levels of change as possible into account, we have to consider a wide range of possible instantiations of one phraseme—preferably automatically formulated as search string—to be looked up in a corpus.

3 Resources

3.1 Corpora

In recent years there have been several attempts to create diachronic corpora for German for various research perspectives. Given our interest in the time from 1650 until today, two corpora are specifically relevant: *Deutsches Textarchiv* (= DTA)[1] with 532 texts from 1650 to 1900 and *GerManC* with texts from 1650 to 1800 (Bennett et al., 2010). DTA aims to make available the most relevant cross-disciplinary German-language books. GerManC aims to provide "a basis for comparative studies of the development of the grammar and vocabulary of [...] German and the way in which they were standardized."[2] The corpus consists of representative 2000-word samples from nine genres from various regions.

While DTA makes available whole texts under a Creative Commons License, GerManC provides snippets only. However, both projects aimed to digitize the most authentic

[1] http://www.deutschestextarchiv.de
[2] http://www.llc.manchester.ac.uk/research/projects/germanc

versions of the texts, i.e., first editions. There are other freely available collections of texts from the relevant time, like the online library provided by TextGrid—also under a Creative Commons License—with texts from the beginning of publishing until the beginning of the 20th century[3]. This collection is volume 125 from the "Digitale Bibliothek" (DB125), consisting of roughly 2700 fictional texts with about 87 million running word forms. These texts are usually later editions.

We will also be able to explore two special-purpose collections: *Text+Berg digital* (Volk et al., 2010) consisting of the yearbooks of the Swiss Alpine Club from 1864 to 2009 with 36 million running word forms, and a subset of the *Collection of Swiss Law Sources* (Gschwend, 2007), containing about 4 million running word forms in texts from ca. 1000 to 1798 (Piotrowski, 2010).

3.2 Dictionaries and collections

Synchronic phraseological information at various points in time can be found in general-purpose dictionaries like Adelung (1801) (= Adelung), Campe (1812), or Sanders (1865), or in special-purpose dictionaries like Wander (1880) (= DSL), Friedrich (1976), or Dudenredaktion (2008) (= Duden11). These dictionaries list phrasemes known and used at a specific point in time; some, like (Grimm and Grimm, 1971) (= DWB) or Borchardt (1888) integrate some diachronic and etymological information to different extents. Etymological information provided by the folklorist Röhrich (2002) often tends to be quite vague, using expressions like "originally" or "formerly". Looking at one dictionary or collection at a time, we get mainly synchronic impressions; moreover, the sources are rarely, if ever, identified, and of course every dictionary claims to be the most authoritative one.

If a phraseme in Duden11 is not listed in DSL, like *einen Quantensprung machen* ("to make a quantum leap", 'to make huge progress'), this might be evidence that this specific phraseme was not known 130 years ago, possibly (as in this case) because one of the lexical units or the concept was not known then. Vice versa, if a phraseme is listed in an older dictionary, but not in a modern one, this might be evidence that this phraseme is not used any more, like *einen Krebs im Beutel haben* ("to have a crab in the bag", 'to be short of money').

To get a first impression of diachronic change, we explored existing dictionaries and collections from different dates. Some, like DSL were already available in electronic format, others, like Duden11 were digitized by us for internal use.[4] Comparing listed phrasemes allows phraseologists to decide which phrasemes to inspect further because of (potential) changes in use, meaning, and/or pragmatics. After analyzing the dictionary data and selecting phrasemes, we will search for evidence in the corpora described in section 3.1 and annotate the results to serve as source for describing diachronic changes.

Fischer and Keil (1996) distinguish non-compositional and compositional idioms, referring to syntactic and semantic flexibility, the latter allowing to vary parts of

[3] `http://www.textgrid.de/digitale-bibliothek.html`

[4] In the meantime, Duden11 is available online.

the phraseme by adjectival modifications, quantification, or by using demonstrative determiners. We do not follow this differentiation, and allow for variation in all phrasemes of interest. There is also no consensus among phraseologists on how to define and to distinguish *variants* and *synonyms*. We therefore provide information about the characteristics of similarity of two examples concerning form and meaning. This allows for searching and displaying phrasemes (or their instantiations in our corpus) by formal aspects—e.g., sharing similar syntactic structure or lexical units—or by meaning—e.g., expressing similar semantic concepts.

4 Challenges

Given our aim and considering the limited resources with respect to manpower and time, we face several challenges.

First, it is impossible to investigate *all* phrasemes listed in one or more of the available collections: we were able to extract 33,200 potential phrasemes from Küpper (1997) (=WdDU), 45,729 potential phrasemes from DSL, 11,500 from "Redensartenindex" (RA-I)[5], 13,300 from Duden11, etc. Due to different conventions for formulating the citation form used by the authors or editors of these collections and dictionaries, it is impossible to compare the entries as such to detect phrasemes listed in more than one collection. However, even given the smallest number of extracted phrasemes—3'834 manually annotated phrasemes from Adelung—it was too much to explore the diachronic evolution of all of them. We therefore had to select phrasemes meeting several constraints: they should be of interest for the intended audience (i.e., researchers as well as laypersons), the phrasemes should have undergone some change over time, the sample should not be restricted to the most common or most unknown phrasemes used today, and there should be a certain frequency of occurences of the phrasemes in corpora. In section 5.1 we report on this aspect.

Second, the corpora of interest for us do not come in comparable formats. All corpora mentioned in section 3.1 are annotated according to the TEI guidelines. TEI P5 (Wittern et al., 2009) allows projects to use various subsets of TEI, so that actual annotation may differ between corpora; however, most of the corpora are not deeply annotated, we can reduce all annotation to the most shallow one, allowing for an easy mapping between annotations to provide a common ground. What is more important, although for example DTA put a lot of effort in normalizing and lemmatizing the texts (with very good results, see for example Jurish (2010)), users can download the non-lemmatized texts only. The Collection of Swiss Law Sources also provides no normalization or lemmatization. The same is true for the TextGrid library. Only the small corpus of alpine texts provides lemmatized texts. As we cannot investigate normalization or lemmatization of old German language variants during project time and as today there are no such tools available providing reasonable quality to be applied without further

[5] `http://www.redensartenindex.de`

effort, we cannot use standard corpus-linguistic tools as in the Idiomdatenbank project. In section 5.2 we report on first attempts to overcome this issue.

Third, searching for phrasemes in corpora means looking for evidence of a sequence of words allowing for inclusion of particles or adjectives as well as for morphological and syntactical variation. Whether a found sequence is indeed a phraseme or whether the words are used literally, can only be decided by looking at the context; in most cases this decision has to be made by the phraseologist, who generally cannot rely on intuition if it comes to older texts. It is hardly possible to reduce this manual effort (see also Rothkegel (2007)).

5 Approach

From a diachronic point of view, polylexicality involves language changes on various levels—structure and meaning of the lexical components as well as structure and meaning of the whole phraseme. When deciding on which phrasemes to investigate we have to allow for changes on all levels to find evidence for these phrasemes in our corpora. As mentioned in section 4, we have to define a sample of phrasemes used for investigation. The OLdPhras dictionary will contain entries with detailed descriptions of their development, while for others we will provide selected information only. We will first report on how to select this sample and we then develop searching strategies for our corpora.

5.1 Choosing the sample of phrasemes for investigation

From Adelung and DSL (both in digitized versions) we extracted phrasemes representing German in the late 18th and in the 19th century—the *historical phrasemes* (HP). For DSL we could make use of typographical structuring of the entries and extract all potential phrasemes automatically. As there was no such typographical structure used in Adelung, the text was annotated manually using the author's markers as starting point[6]—which had the advantage that information concerning meaning, use, and variation could be annotated as belonging to a specific phraseme at the same time. We used *Stripey Zebra*, the current version of the German Malaga Morphology (Lorenz, 1996) to identify nouns used in the extracted phrasemes.[7] The continuously most frequent nouns indicate a constant productivity of components: for somatisms, i.e., words for body parts, like hand, head, eye, ear, or nose, we find a great number of phrasemes.[8]

[6] Adelung used markers like "Sprichwort", "Redensart" or "RA", but not consistently.

[7] *Stripey Zebra* is a rule-based morphological analyzer, providing detailed, hierarchically structured results; using pruning and weighting *Stripey Zebra* can provide "the best" analysis according to the morphological principles of derivation, compounding, and inflection. For unknown words a hypothesis is generated. See Mahlow and Piotrowski (2009) for a detailed description and performance data.

[8] Using a modern morphological analyzer like *Stripey Zebra* poses some bias, as older spelling variants might result in wrong results or no results at all. However, manual inspection showed

Extracting potential phrasemes from contemporary sources—Duden11 and RA-I, the *contemporary phrasemes* (CP)—by making use of typographical information and using lemmatization again, we could compare rankings of nouns. We were interested in nouns being part of a large variety of phrasemes today *and* in former times—suggesting ongoing productivity (we did not compare the individual phrasemes in which they occur, but only the number of phrasemes). Focusing on changes, we were also interested in nouns showing higher productivity in older collections than in newer ones and vice versa—indicating loss of phrasemes and emergence of new ones.

We set a threshold of 2%, meaning that a noun was considered *frequent*, if it belongs to the top 2% in the frequency list of all nouns of a collection. A noun was considered *infrequent*, if it was found in one or two phrasemes of a collection only[9]. Based on this we assembled (a) historical frequent nouns, (b) contemporary frequent noun, (c) historical infrequent nouns, and (d) contemporary infrequent nouns.
Based on that we could identify:

- Nouns frequently used in HP *and* in CP, like *Hand* ("hand"), *Teufel* ("devil"), or *Kopf* ("head")
- Nouns frequently used in HP, but not in CP, especially animals like *Affe* ("monkey"), *Laus* ("louse"), or *Kuh* ("cow"), as well as *Narr* ("fool"), *Schnee* ("snow"), or *Feder* ("feather")
- Nouns frequently used in CP, but not in HP, *Nerv* ("nerve"), *Fall* ("case"), or *Punkt* ("point")
- Nouns infrequently used in HP *and* in CP, like *Affenschande* ("apish shame"), *Friedenspfeife* ("calumet"), or *Gnadenbrot* ("charity")
- Nouns infrequently used in CP, but more frequently in HP, like *Krebs* ("crab"), *Käse* ("cheese"), or *Weib* ("woman")
- Nouns infrequently used in CP, but not used at all in HP, like *Fleischwolf* ("meat grinder"), *Brechstange* ("crow bar"), *Sprungbrett* ("diving board"), or *Abstellgleis* ("holding track")

Keeping interested laypersons in mind, we did not look for infrequently used nouns in HP with no evidence in CP or more frequently used in CP.

Having identified diachronically "interesting" nouns, we explored the phrasemes in which these nouns occur. For each resource we independently identified and allocated variants and synonyms of phrasemes by assigning specific *phraseme types*. A phraseme type represents a specific semantic concept. Instantiations include all lexical and

that results on the nouns occuring in our extracted phrasemes, are quite acceptable, there is more spelling variation in verbs.

[9] We decided to use two phrasemes as the lower bound instead of one, as manual inspection had shown that for nouns with two associated phrasemes, the phrasemes typically tend to be variants of each other.

structural variants. For example, in Adelung we find the phraseme *jemanden Staub in die Augen streuen* ("to throw dust into someone's eyes", 'to pull the wool over someone's eyes'), whereas in Duden11 we have *jmdm. Sand in die Augen streuen* ("to throw sand into someone's eyes"), both of them belong to the same phraseme type and express the same meaning.

Based on phraseme types, we can then flip the matrix and see for each phraseme type which nouns in which phrasemes are associated to this particular phraseme type in which collection. We thus get an impression of the variants already reported in various collections and can thus decide which phraseme types to investigate further. By annotating other resources like Borchardt (1888) or WdDU with phraseme types as well, we create a rich resource that allows us to get a first diachronic impression. Note that up to this moment, we have made use of already existing lexical information, which we have rearranged and recombined. We still lack empirical evidence but rely on statements of other phraseologists only. In the next section we look at the empirical part, which is work in progress.

5.2 Searching corpora

With respect to corpora, we have to face a different notion of *frequent* and *infrequent*: an infrequent noun like *Friedenspfeife* with only one associated phraseme type (*die Friedenspfeife mit jemandem rauchen* "to smoke the calumet together", 'to reconcile') might be found quite often in texts and thus be relatively frequent. Additionally, we can calculate the frequency of the lexical units of a phraseme as occurring in the text regardless of whether it is used in a phraseme or in its literal meaning in other contexts. However, one fundamental problem searching corpora for phrasemes is their generally low frequency compared to other multi-word units. (Colson, 2007) Due to variation of the phrasemes and to the decreasing size of corpora, phrasemes get more and more difficult to find the further back in time we search (see also Claridge, 2008).

Our first intuition—based on our definition of variants and synonyms of phrasemes and taking into account the state of the art—was a lemma-based search for phrasemes and their variants, allowing for syntactical, lexical, and morphological variation. However, as most of the relevant corpora do not provide lemmata and we will not be able to lemmatize them automatically, we have to come up with other strategies, taking into account spelling variants, too.

Using vector-based approaches from the field of information retrieval (IR) (Salton et al., 1975) like computing co-occurence vectors for the phrasemes in question, we can use the already identified instantiations of a phraseme type to find them in the corpora by matching the query-vector to the corpus allowing for variation—the phraseologists will then have to decide if a match is indeed idiomatic. However, we also have to take into account that the texts in the corpora are written in several variants of German.

Spelling variation and different inflectional paradigms[10] might influence recall and precision of vector-based approaches.

For queries considering spelling variation, we will make use of data provided by the project "Freiburger Anthologie", a collection of the 1000 most important German poems.[11] We have enriched this data with observations in the texts of DB125. We found further spelling variants, out-dated vocabulary, and different inflectional paradigms especially for verbs. Note that using vectors we can look for surface-similarity only, not for semantic similarity.

For finding variants including semantically similarity we will use GERMANET (Hamp and Feldweg, 1997) to identify synonyms, hyperonyms, and hyponyms for the lexical units used in a phraseme. For example from *der Apfel fällt nicht weit vom Stamm* ("the apple does not fall far from the stem", 'like father like son') we can create the forms *der Apfel fällt nicht weit vom Baum* ("the apple does not fall far from the tree"), *die Birne fällt nicht weit vom Stamm* ("the pear does not fall far from the stem"), *die Birne fällt nicht weit vom Baum* ("the pear does not fall far from the tree"). Considering spelling variation and allowing for changes in word order we are then able to find *die birn nit wey vom baum falt* (Rechtsquellenstiftung des Schweizerischen Juristenverbandes, 2009, 121)

Automatically creating search queries including semantic variation for single lexical units will allow us to automatically create variants of the phrasemes we are investigating. Using vector-based IR algorithms we will then look for evidence in the corpora. The results will contain context to enable phraseologists to decide whether a particular match is a phraseme or a non-idiomatic co-occurrence only. If a match is not a phraseme, but the words are used literally, the match will not be rejected but marked as non-idiomatic. Idiomatic evidence will be annotated with all information needed to serve as source for a general comment on diachronic change of the phraseme under investigation as well as information to be directly displayed to the user of the resulting dictionary.

Based on the results of all these procedures described above, we will be able to enrich the lexicographic information for a particular phraseme type. We will also be able to provide statistical information showing some trends concerning increase, decrease, or stability of use of a phraseme type (or a specific variant) over time.

6 Conclusion

We presented our semi-automatic approach for investigating phrasemes in German from a diachronic perspective. Due to the diachronic aspect, several issues arise which can be solved by using manual effort in combination with automatic processing steps. Searching for variants of phrasemes (or multi-word units in general) in historical texts

[10] A word might have belonged to a different inflectional paradigm a few hundred years ago than it does today, an example would be the verb *backen* ("to bake") with weak inflection today (*backte* and strong inflection formerly *buk*. However, the strong inflection is still used in Swiss Standard German, but not in Germany or Austria.

[11] `http://freiburger-anthologie.ub.uni-freiburg.de/fa/fa.pl?cmd=gedichte\&sub=analog\&add=`

emphasizes the need to solve problems of normalization and lemmatization—higher-level applications as the OLdPhras project rely on those annotations to allow the use of state-of-the-art NLP, information retrieval, or text mining tools. In particular, if lemmatization has already been performed, freely available corpora should be distributed including this annotation.

However, including the human in the loop at various steps at the process, we developed a semi-automatic approach that is transferable to other situations—other languages or texts from other periods. We will thus be able to provide some information on the evolution of form, meaning, and use of German phrasemes that goes beyond example-based explorations. We will also be able to annotate respective information in the corpora, which might later be used by other researchers investigating other questions.

7 Acknowledgments

We thank our colleagues from the OLdPhras project for collaboration on concepts and for the thorough manual annotation of the various extracts described. We also thank the anonymous reviewers for helpful comments on an earlier version of this paper.

References

Adelung, J. C. (1793–1801). *Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart*. Breitkopf & Sohn, Leipzig. (= Adelung).

Bennett, P., Durrell, M., Scheible, S., and Whitt, R. J. (2010). Annotating a historical corpus of German: A case study. In *Proceedings of the LREC 2010 Workshop on Language Resource and Language Technology: Standards - state of the art, emerging needs, and future developments*, pages 64–68, Paris. ELRA.

Borchardt, W. (1888). *Die Sprichwörtlichen Redensarten im deutschen Volksmund nach Sinn und Ursprung erläutert*. Brockhaus, Leipzig.

Burger, H. (2010). *Phraseologie*. Erich Schmidt, Berlin.

Burger, H. and Linke, A. (1998). Historische Phraseologie. In Besch, W., Betten, A., Reichmann, O., and Sonderegger, S., editors, *Sprachgeschichte: Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*, pages 743–755. Walter de Gruyter, Berlin/New York.

Campe, J. H. (1807–1812). *Wörterbuch der deutschen Sprache*. Schulbuchverlag, Braunschweig.

Claridge, C. (2008). Historical corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics*, pages 242–259. Walter de Gruyter, Berlin/New York.

Colson, J.-P. (2007). The World Wide Web as a corpus for set phrases. In Burger, H., Dobrovol'skij, D., Kühn, P., and Norrick, N. R., editors, *Phraseology*, Handbooks of Linguistics and Communication Science, pages 1071–1077. Walter de Gruyter, Berlin/New York.

Dudenredaktion (2008). *Redewendungen: Wörterbuch der deutschen Idiomatik*. Dudenverlag, Mannheim. (= Duden11).

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. PhD thesis, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung.

Fellbaum, C., editor (2007). *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies*. Research in Corpus And Discourse. Continuum, London/New York.

Fellbaum, C. and Geyken, A. (2005). Transforming a corpus into a lexical resource the Berlin Idiom Project. *Revue française de linguistique appliquée*, X(2):49–62.

Fischer, I. and Keil, M. (1996). Parsing decomposable idioms. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 388–393, Stroudsburg, PA, USA. Association for Computational Linguistics.

Friedrich, W. (1976). *Moderne deutsche Idiomatik. Systematisches Wörterbuch mit Definitionen und Beispielen*. Huber, München.

Fritzinger, F., Kisselew, M., Heid, U., Madsack, A., and Schmid, H. (2009). Werkzeuge zur Extraktion von signifikanten Wortpaaren als Webservice. In Hoeppner, W., editor, *GSCL-Symposium Sprachtechnologie und eHumanities*, pages 32–43.

Grimm, J. and Grimm, W. (1852–1971). *Das deutsche Wörterbuch*. Hirzel, Leipzig. (= DWB).

Gschwend, L. (2007). Die Sammlung Schweizerischer Rechtsquellen, herausgegeben von der Rechtsquellenstiftung des Schweizerischen Juristenvereins: Ein Monumentalwerk rechtshistorischer Grundlagenforschung. *Zeitschrift für Schweizerisches Recht*, 126(I):435–457.

Hamp, B. and Feldweg, H. (1997). GermaNet – a lexical-semantic net for German. In *Proceedings of the ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Somerset, NJ, USA. Association for Computational Linguistics.

Heid, U. (2007). Computational linguistic aspects of phraseology II. In Burger, H., Dobrovol'skij, D., Kühn, P., and Norrick, N. R., editors, *Phraseology*, Handbooks of Linguistics and Communication Science, pages 1036–1044. Walter de Gruyter, Berlin/New York.

Jurish, B. (2010). More than words: Using token context to improve canonicalization of historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–39.

Juska-Bacher, B. and Mahlow, C. (2012). Phraseological change – a book with seven seals? tracing back diachronic development of German proverbs and idioms. In Durell, M., Scheible, S., and Whitt, R. J., editors, *TBA*, volume 3 of *Corpus linguistics and Interdisciplinary perspectives on language*. Gunter Narr, Tübingen, Germany.

Keil, M. (1997). *Wort für Wort – Repräsentation und Verarbeitung verbaler Phraseologismen (Phraseo-Lex)*, volume 35. Max Niemeyer Verlag, Tübingen.

Kluge, F. (2002). *Etymologisches Wörterbuch der deutschen Sprache*. Walter de Gruyter, Berlin, New York, 24th revised and expanded ed edition.

Kühn, P. (2003). Phraseme im Lexikographie-Check: Erfassung und Beschreibung von Phrasemen im einsprachigen Lernerwörterbuch. *Lexicographica*, 19:97–118.

Kuiper, K., McCann, H., Quinn, H., Aitchison, T., and van der Veer, K. (2003). *SAID: A syntactically annotated idiom database*. Linguistic Data Consortium, Philadelphia.

Küpper, H. (1997). *PONS Wörterbuch der Deutschen Umgangssprache*. Klett Verlag, Stuttgart. (= WdDU).

Lorenz, O. (1996). Automatische Wortformerkennung für das Deutsche im Rahmen von MALAGA. Master's thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Mahlow, C. and Piotrowski, M. (2009). A target-driven evaluation of morphological components for German. In Clematide, S., Klenner, M., and Volk, M., editors, *Searching Answers – Festschrift in Honour of Michael Hess on the Occasion of his 60th Birthday*, pages 85–99. MV-Verlag, Münster.

Mellado Blanco, C. (2009). Einführung. Idiomatische Wörterbücher und Metafraseografie: zwei Realitäten, eine Herausforderung. In Mellado Blanco, C., editor, *Theorie und Praxis der idiomatischen Wörterbücher*, pages 1–20. Max Niemeyer, Tübingen.

Piotrowski, M. (2010). From Law Sources to Language Resources. In Sporleder, C. and Zervanou, K., editors, *Proceedings of the ECAI 2010 workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, pages 67–71.

Rechtsquellenstiftung des Schweizerischen Juristenverbandes, editor (2009). *Appenzeller Landbücher*, volume SSRQ AR/AI 1 of *Sammlung Schweizerischer Rechtsquellen*. Schwabe, Basel, Switzerland.

Röhrich, L. (2002). *Das große Lexikon der sprichwörtlichen Redensarten*. WBG, Darmstadt.

Rothkegel, A. (2007). Computerlinguistische Aspekte der Phraseme I. In Burger, H., Dobrovol'skij, D., Kühn, P., and Norrick, N. R., editors, *Phraseology*, Handbooks of Linguistics and Communication Science, pages 1027–1035. Walter de Gruyter, Berlin/New York.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sanders, D. (1859–1865). *Wörterbuch der deutschen Sprache*. Wigand, Leipzig.

Seretan, V. and Wehrli, E. (2010). Tools for syntactic concordancing. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 493–500. IEEE.

Stantcheva, D. (2003). *Phraseologismen in deutschen Wörterbüchern: Ein Beitrag zur Geschichte der lexikographischen Behandlung von Phraseologismen im allgemeinen einsprachigen Wörterbuch von Adelung bis zur Gegenwart*. Dr. Kovač, Hamburg, Germany.

Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., and Ruef, B. (2010). Challenges in building a multilingual Alpine heritage corpus. In *Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 1653–1659, Paris. European Language Resources Association (ELRA).

Wander, K. F. W. (1867–1880). *Deutsches Sprichwörter-Lexikon*. Brockhaus, Leipzig. (= DSL).

Wittern, C., Ciula, A., and Tuohy, C. (2009). The making of TEI P5. *Literary and Linguistic Computing*, 24(3):281–296.