Transforming a Corpus into a Lexical Resource–The Berlin Idiom Project

Christiane Fellbaum
Department of Psychology, Princeton University
Princeton, New Jersey, USA
Berlin-Brandenburg Academy of Sciences
fellbaum@princeton.edu
and
Alexander Geyken
Berlin-Brandenburg Academy of Sciences
Berlin, Germany
geyken@bbaw.de

July 15, 2005

Abstract

We discuss the goals and methods of the lexicographic project "Collocations and Idioms in the German Language" at the Berlin-Brandenburg Academy of Sciences. A very large corpus is tagged and parsed to enable flexible searches for target structures, German verb phrase idioms. On the basis of relevant tokens, an extensive linguist-lexicographic analysis is performed and recorded on a set of structured forms, which comprise a kind of digital dictionary entry for the target structure. For transparency and future research, each recorded linguistic-lexicographic phenomenon is linked with appropriate corpus tokens. The resulting resource, which combines an exhaustive description of the idioms' properties with corpus tokens, allows for multiple search types.

1 Introduction

1.1 Idioms as fixed lexical items in grammar

For a long time, idioms and collocations were assigned no special status in the lexicon but were considered merely as long words, units of meaning whose forms happen to include several words. The standard example, English *kick the bucket* and its equivalents in other languages, lent itself well to analyses arguing for the non-compositionality and syntactic frozenness of idioms. Nunberg at al. (1994), Dobrovolskij (1999), Moon (1998), Fleischer (1997) and others showed that not all idioms are alike but vary with respect to semantic opacity and syntactic flexibility. Moreover, Nunberg et al. first proposed that these two characteristic properties are correlated: speakers treat those idiom constituents to which they assign a meaning as syntactially free, similarly to literally interpreted constituents. Psycholinguists such as Glucksberg (1993) also found that noun constituents that have metaphoric status are available for syntactic operations. Few researchers still cling to the view that all idioms are semantically non-compositional, syntactically frozen "long words," and there is general agreement that idioms are distributed along a scale of fixedness and variability. However, the grammar of both individual idioms and possible classes of idioms has not been sufficiently explored, largely due to a lack of empirical data. More generally, the syntactic and lexical variability of idioms raises the question as to their place in the grammar, as their status as lexical units is anything but straightforward. The absence of large amounts of corpus data necessarily limited linguistic inquiry to constructed data and the intuition of the working linguist. In the past decade large corpora have become available and empirical investigations are now the *sine qua non* for theoretical work. Corpora clearly show the limitations of data produced out of context by an individual speaker. This is particularly true of idioms, which lend themselves to playful variations conditioned by specific contexts. A wide range of morphosyntactic operations and lexical variations show that most idioms are more deeply integrated into the grammar than is often assumed.

1.2 Idioms in traditional dictionaries

Traditional lexicography largely has not done justice to idioms, either. This is partly due to the lack of empirical investigations into the variations that idioms undergo in actual use.

Moreover, lexical resources tend to be bound by the constraints imposed by their format. By virtue of being "word books," most lexicons represent

idioms on par with words, contributing to the perception of these phrases as long words. The grammatical information that would adequately account for an idiom's uses falls outside the scope of a lexicon. Striving to be helpful to dictionary users, native speakers and learners alike, lexicographers often represent the idioms in a context that seems typical or frequent. But corpora show a surprising variety of uses, and often the "typical" cases are in fact less frequent than believed. Writers of traditional dictionaries are usually mindful of non-native users. When such speakers create playful, context-specific variations of idioms, they risk being misunderstood and corrected rather than commended for their command of the language. As a result, didactic materials and dictionaries aimed at learners tend to present idioms in what are considered the most frequent or typical uses.[1]

The work described here is carried out within the framework of a long-term corpuslinguistic and lexicographic project that aims to give a broad, in-depth account of the use of German idioms. In the remainder of the paper, we (a) outline our method for building a lexical resource for German idioms, (b) describe the corpus that forms the empirical basis of the investigation and the methods for extracting candidate strings, (c) discuss specific examples of the linguistic and lexicographic phenomena that are recorded, and (d) sketch the resultant database combining corpus examples with rich lexicographic-linguistic annotations.

Our focus are Verb Phrase (VP) idioms, i.e., verb phrases with a verbal head and at least one complement (NP or PP). We exclude particle verbs as well as support and light verb constructions from the present analysis.

2 Methodology: Overview

We identified some 2,000 VP idioms as candidate target structures. The choice was guided by the frequency of the verb heads; the list includes high-frequency verbs like *machen* (make), *bringen* (bring), and *stehen* (stand).

Search queries are written to extract all and only the appropriate strings from the corpus (cf. Section 3). Retrieved tokens are manually sorted into idiomatic uses, non-idiomatic uses, and non-matches. (The corpora of contexts for idiomatic and non-idiomatic uses of polysemous strings could be valuable for training automatic systems to learn to recognize and distinguish the intended meanings.) The idiomatic uses of a given target string are collected into an example corpus.

[1] We thank Karin Aguado for a helpful discussion on this point.

In the next step, a full linguistic-lexicographic analysis of the example corpus data is performed and recorded on a "template" (cf. Section 4). Each phenomenon of interest is linked to the appropriate corpus examples, ensuring full transparency of the analyses. The resultant resource combines sophisticated lexicographic entries with corpus data.

3 Corpus, extraction of search patterns

3.1 Corpus

Over the course of two and a half years (2000 and 2003) the project *Digitales Wörterbuch der Deutschen Sprache* (Digital Dictionary of the German Language) at the Berlin-Brandenburg Academy of Sciences created two different corpora of the 20th century German language: the balanced DWDS-Kerncorpus (core corpus) and the opportunistic DWDS-Ergänzungscorpus (supplementary corpus) (Geyken and Klein, 2004). Together, they form the basis for the investigation of German idioms.

The DWDS-Kerncorpus consists of 100 million tokens, thus matching in size the Britisch National Corpus and the American National Corpus currently under construction. It is a balanced corpus of German texts of the 20th century, i.e. it is equally distributed over time and over the following five text types: journalism (approx. 27% of the corpus), literary texts (26%), scientific literature (approx. 22%) and other non-fiction (approx. 20%), as well as transcripts of spoken language (5%). This classification has mostly practical reasons. As all texts are encoded following the TEI-guidelines, it would be comparatively easy to introduce more fine-graded text types or to classify the texts according to a different classification code.

Besides journalistic texts (newspaper reports and articles from periodicals taken from more than 50 different newspapers and magazines), there are literary monographs, poetry and dramatic works. Non-fiction texts such as cook books, maintenance manuals, and guides to etiquette are found in the corpus as well as important scientific works (e.g. Einstein, R. Koch, J. Habermas).

The much larger DWDS-Ergänzungscorpus contains around 900 million words of running text, mainly from daily and weekly newspapers of the 1990s such as *konkret*, *Frankfurter Allgemeine Zeitung*, *Frankfurter Rundschau*, *Neue Zürcher Zeitung*, *Spiegel*, *taz*, and *Die ZEIT*. More than 2 million newspaper articles have been gathered in this corpus.

3.2 Extraction of idioms

DDC - a linguistic search engine The basic tool for searching the DWDS-corpus is the linguistic search engine DDC (Dialing DWDS Concordancer) (Sokirko 2003). DDC was developed in Moscow and optimized in Berlin. DDC is a search engine developed specially to meet the needs of linguistic queries. Input queries may consist of sequences of word forms, lexical categories, lemmata, thesaurus elements, or combinations of all four. Also supported are right and left-truncated searches, Boolean AND, OR, NOT searches, and interval searches (NEAR, FOLLOWED_BY), and regular expressions (except for negation).

DDC itself is only a search engine. In order to allow for linguistic queries, corpus texts have to be annotated with morphological information. This is done via a slightly extended version of Morphy (Lezius 2000, Sokirko 2004). Morphy associates to each token its lemma and part-of-speech category as well as its morphological features.

Extraction of search patterns As the goal of the idiom project is to observe the behavior of idioms in corpora, the query expressions have to be sufficiently general in order to capture all the variations to the basic form of a given idiom such as the lexical substitution of obligatory constituents and the elision or insertion of optional elements. On the other hand, queries must not be too general since – given the high frequency of verb head – one could end up with tens of thousands of hits, a number which is no longer tractable for manual investigation. It is important to bear in mind that we rejected a purely automatic classification of the cases by a syntax parser since verb-noun idioms frequently violate morphological and syntactic regularities, thus making automatic syntax parsing a quite error prone task.

The following examples illustrate some of the difficulties one faces for the flexible extraction of idioms in the DWDS corpus. To date (June 2005), more that 4,000 different queries have been formulated for more than 1,000 verb noun expressions.

Rare lexemes and hapax legomena Verb phrase idioms frequently contain rare or obsolete lexemes or lexemes not used outside the idiom. In such cases it is safe to formulate a query that consists of only such a lexeme rather than a multi-word query. For example, the noun in the idiom *Fersengeld geben* (lit. give heel money: take a powder, vanish) is unique to the idiom. Although this idiom generally occurs with the verb *geben* (give), the query

- `Fersengeld && (geben || Geben)` (84 hits)

is too specific since the result set does not return lexical variations of the verb. However, the analysis of all hits containing the word form *Fersengeld* (102 hits) yields 18 cases of the variation *Fersengeld zahlen* (pay heel money).

Word order Because few idioms occur only in one word order, it is necessary to formulate queries with a symmetric distance to a target lexeme. For example, for the idiom *auf der Lauer liegen* (lie in wait) the straightforward asymmetric query *Lauer fby 5 liegen*, i.e. the noun *Lauer* is followed by the verb in a distance of at most 5 tokens, would miss cases with relativization such as *es gibt keine Lauer, auf der ich liege* (there is no wait that I don't lie in). Hence, we generally formulate the symmetric query `NEAR(Lauer,liegen,5)`.

Optional elements One major difficulty for the appropriate idiom extraction is to stipulate which parts of the idiom are obligatory and, conversely, which parts can be omitted or substituted with other elements. Here, introspection of native speakers can sometimes be an obstacle. For example, native speakers differ in their judgement with respect to the preposition *aus/auf dem letzten Loch pfeifen* (lit., blow from/on one's last hole: be on one's last leg). Speakers who accept *aus* generally reject the variant with *auf* and vice versa. In fact, the two variants are distributed almost equally throughout the DWDS corpus: there are 64 occurrences with *aus* and 80 with *auf*. Hence, it is necessary to vary the different components of the idiom with some heuristics and verify manually whether these tests yield recurrent patterns that can be related to the idiom.

Distance and idiomaticity In a vast corpus like the one-billion DWDS corpus, it is not feasible to manually verify all the extracted co-occurrences of idiom components with all its stipulated variations, as one would end up with tens of thousands of hits. In order to reduce the number of sentences submitted to manual analysis we try to pre-classify the different co-occurrences. It turned out that the distribution of the frequencies of two co-occurring idiom components according to their distance gives some insights into the syntactic structure of a given idiom (Herold 2005a, 2005b). For example, the two idiomatic expressions *Rede halten* (make a speech) and *Gefahr laufen* (run the danger (of)) show different graphs as to their distance-frequency-distribution (cf. Figs. 1 and 2 below). *Rede halten* is almost symmetric and the number of idiomatic occurrences decreases slowly

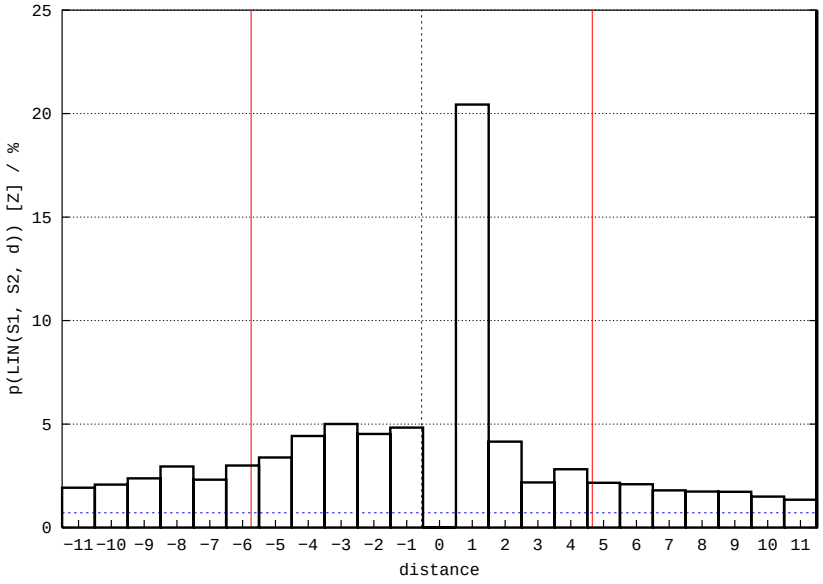

Figure 1: distance-frequency-distribution of *Rede halten*

with an increasing distance. By contrast, *Gefahr laufen* exhibits a strong bias to the negative distance (i.e., the noun is followed by the verb); this idiom occurs with significant frequency only for the distances of -4 to +2. These differences in frequency reflect the finding that the percentage of true idioms in the result set is much higher for *Gefahr laufen* than for *Rede halten*.

The classification of queries according to their distance has consequences for the manual analysis of idioms. For example, it might be more interesting to look at corpus patterns of different distances rather than to go through thousands of recurring syntactic patterns of a certain distance. This is the case for *Rede halten*, where 2,305 out of 11,278 occurrences correspond to the adjacent pattern *Rede halten* [2].

[2] with variations only in verb inflections

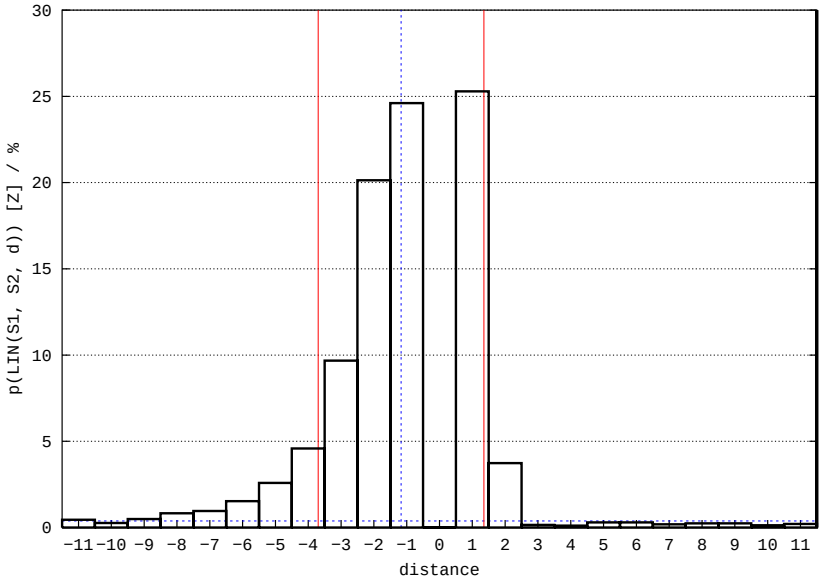

Figure 2: distance-frequency-distribution of *Gefahr laufen*

4 The template

Our goal is to create a linguistic-lexicographic resource that represents the uses of hundreds of common German VP idioms in a comprehensive way. The contents of this resource are grounded in corpus data. The data show that, while idioms vary in the uses to which speakers put them, they are not frozen lexical units that are simply "plugged" into larger phrases. They are fully integrated into the grammar and are accessed on all levels of grammar. We are mindful of representing the idioms' grammatical properties in a fairly theory-neutral way so as to make the results accessible and acceptable to future researchers.

We developed a digital annotation scheme or questionnaire, which we dubbed the "template." It serves as both input and output interface and links the annotated example corpora with the idiom knowledge base (Neumann et al., 2004, Kramer et al., 2004). The data entry interface supports a structured entry created by linguists/lexicographers who record an idiom's properties on the basis of the corpus examples. The template design reflects linguistic and lexicographic aspects, focusing on the interplay between "normal" and usage and variation.

The template is structured according to the standard levels of lexico-

graphic and linguistic analysis. In the following sections, we discuss specific examples for each.

5 Levels of analysis

5.1 Citation forms and norms

In order to determine what constitutes a variation, one must necessarily proceed from a norm. Traditional dictionaries, being organized alphabetically, are organized around lemmata, "base forms" that constitutes some kind of unmarked norm. For single-word lexemes, this is uncontroversially the uninflected singular form of a noun and the infinitive of a verb. Idiom dictionaries like Schemann (1993) list most VP idioms with an infinitival active form and its complements. But inspections of hundreds of corpus tokens of an idiom show that the notion of a "norm" is not at all given.

Our criterion for determining an idiom's citation form for the template is frequency. However, in some cases, the corpus tokens are evenly divided two or three different forms, necessitating multiple citation forms and templates.

The citation forms often are not morphologically unmarked. The verbs in certain idioms occur preferentially in a specific person or voice. Another difficulty is determining the status of the constitutents – are they obligatory or optional components of the idioms? The richness of the variations sometimes makes this question difficult to answer. The template provides for "core constituents," "optional constituents," and "idiom-external constituents" (the latter are frequently the subjects of the VPs).

Some idioms occur characteristically with a limited set of semantically specific adverbs or other external modifiers. Such frequent but non-obligatory components are labelled in the template as "preferred syntactic environment."

5.2 Phonology

Certain idiomatic variations turn on the phonological property of the idiom's "base" form. This can be clearly seen in the following examples, where semantically unrelated homophonous words have been inserted into the idioms:

(1) ein letzter, weit über die *Stränge* (und die *Strenge* des
a last, far beyond the lines (and the strictness of the

nun beginnenden Berufslebens) schlagender Abschied
now beginning working life) going evening

an evening that exceeded the norms (and the strictness of the working life that would now begin)

(2) Er hoffe nun, nicht *Spießrouten* durch die Anlage laufen zu müssen
he hopes now, not the gauntlet/routes through the place to run to have
He is now hoping not to have to run the gauntlet/routes through the place
(Here, *Spießrouten* is substituted for *Spießruten*, gauntlet)

A related case is illustrated in (5), where the the speaker merges two idioms (3) and (4), inserting two phonologically similar NPs from one idiom into another:

(3) Hinz und Kunz
Hinz and Kunz
Tom, Dick, and Harry

(4) von Pontius zu Pilatus laufen
from Pontius to Pilate run
run from pillar to post

(5) von Hinz zu Kunz laufen

This is a classic example of contamination, where the speaker merges two idioms with similar phonological and morphological patterns.

5.3 Morphology

The constituents of many idioms show regular alternations and inflectional paradigms consistent with grammar. Here are some examples:

(6) ..über den Strang schlagen (citation form)
...strike out beyond the rope (go too far)
einen über **alle Stränge** geschlagenen Zweikampf
...all ropes (pluralization)

(7) ein Haar in der Suppe finden (citation form)
find a hair in the soup (find s.th. to criticize)

jedes Haar/tausend Haare in der Suppe
each/a thousand hairs..(determiner variation)

(8) den Bock zum Gärtner machen (citation form)
turn the buck into the gardener (put an incompetent person in charge)
die **Ziege** zur **Gärtnerin** machen
turn the goat into the (female) gardener (lexical and morphological gender variation)

(9) X hat es faustdick hinter den Ohren
X has it fist-thick behind the ears (X is a sly person)
er hat es noch **faustdicker** hinter den Ohren
even fist-thicker (graded adjective)

5.4 Syntax

The syntactic behavior of idioms has long been a subject of discussion among linguists and psycholinguists (Nunberg at al. 1994, Dobrovolskij 1999, and many other). Generally, these studies were based not on corpora but on constructed data. Our corpus shows that many idioms undergo the full range of syntactic transformations. Below are some typical examples.

(10) X nimmt kein Blatt for the Mund
X holds no leaf in front of his mouth (X does not hold back with his opinions)
Ein Blatt nehmen sie dabei vor **keinen** Mund
A leaf they hold before no mouth (Focusing/topicalization with shift of negation)

(11) ein Haar in der Suppe finden
find a hair in the soup (find s.th. to criticize)
Er suchte das Haar in der Suppe, um T. **daraus** einen Strick zu drehen, und schon in der ersten Partie fand er **es**
He looked for the hair in the soup in order to twist a rope for T. from it and already in the first round he found it (relativization)

(12) Was H. gefunden hatte, war nur das Haar in der Suppe
What H. had found was only the hair in the soup (clefting)

(13) Mit einem so kecken Mund, vor den kein Blatt genommen wird
With such a fresh mouth, **in front of which** no leaf is held (relativization)

(14) X nimmt Y auf die Schippe
X takes Y on his shovel (X pulls Y's leg)
...nimmt man alles auf die Schippe, auch **sich selbst**
...one takes everything on one's shovel, even oneself (reflexivization)

(15) irgendwie **findet sich** schon ein Haar in der Suppe
somehow a hair in the soup finds itself (reflexive middle)

5.5 Lexical variation

We also find much variation on the lexical level. It is here in particular that speakers and writers "play" with idioms, creating novel variations that are specific to a context or situation. In such cases, the speaker is fully aware of the idiom's lexical properties; this is not necessarily the case with morphological and syntactic alternations.

(16) X nimmt Y die Butter vom Brot
X takes the butter from Y's sandwich (X steals Y's thunder)
...hat ihm die Argumentbutter vom Wahlkampfbrot genommen
has taken the **argument butter** from his **election campaign sandwich** (compounding)

(17) X hat Y mit der Muttermilch getrunken
X has drunk Y with his mother's milk (X has a natural talent for Y)
...hatte den Fußball schon mit der **Vatermilch** getrunken
had drunk soccer with his **father's milk** (lexical variation of one compound member)

(18) **sexistisch-chauvinistische** Haare in der Suppe (adjectival modification)

Substitution of one of the idiom's components – usually a noun – with another is particularly common. The substituted and the substituting word are typically synonyms, antonyms, meronyms, or other paradigmatically related word forms, and the substituting word relates to the literal reading of the substituted word rather than its meaning in the idiom (if in fact it has one.)

The illustrative data cited here show that many – perhaps most – idioms are by no means frozen structures, but are fully integrated into the grammar. Thus, in so far as idioms belong to the lexicon, they are not fixed structures

but rather meaning units associated with specific lexemes. The presence of one lexeme often suffices to refer to the idiom's meaning, leaving the other lexeme(s) subject to variation. Moreover, corpus data show that the lexemes associated with an idiom need not occur in any syntactictic configuration:[3]

(19) Hier brummt der Bär
The bear is growling here (this is where the action is)
Da ist der Bär im Schilde zu führen, dass es nur so brummt
There are so many plans for the bear that one hears growling

(20) ...auf der ein kuscheliger Braunbär zu sehen ist: Berlin brummt
on which we see a snuggly brown bear: Berlin growls (hums)

Examples like these pose difficulty for arguments that idioms constitute syntactic structures or constructions.

5.6 Variation and semantic opacity

Nunberg et al. (1994) were the first to propose that the syntactic flexibility of idioms is directly correlated with their semantic transparency. Thus, an idiom containing an NP that has metaphorical status and whose meaning can be mapped onto a literal meaning, is expected to be subject to pronominalization, passivization, etc. For example, *beans* in *spill the beans* can be assigned the meaning "secrets" or "confidential information." Consequently, a novel utterance like *he didn't spill a single bean* can easily be interpreted (S. Glucksberg, p.c.). By contrast, *bucket* in *kick the bucket* cannot mapped onto a noun, as "die" is an intransitive verb. Nunberg et al. (1994) predict that *bucket* cannot be passivized, relativized, or modified in any way. However, corpus data show the German equivalent of this idiom, *ins Gras beißen* (lit.: "bite into the grass") being used with internal modification, as in *ins texanische Gras beißen* ("bite into the Texan grass") and *biss ins Gras der weiten Steppen Russlands* ("...bit into the grass of Russia's endless steppes"). Syntactically, the nouns are modified, but in fact the modification seems to have scope over the entire verb phrase, i.e., the event: Someone died in Texas/the endless steppes of Russia.

An even more striking example is the following, where the idiom *den Löffel abgeben* (lit., turn in the spoon, "die") has been broken up, and the

[3]This suggests a resemblance to Chinese "cheng yu," two pairs of characters that express an idiomatic meaning and where one of the pairs, but not both, can be substituted to give the idiom a context-specific reading.

non-referential noun *Löffel* (spoon) is in a semantically and syntactically coordinated structure with *Zepter* (scepter) and *Krone* (crown).

(21) einige Freunde von Prince Charles behaupten, er hoffe...
dass seine Mutter erst dann **die Löffel bzw. Zepter und Krone abgibt**,...
Some friends of Prince Charles claim he is hoping that his mother will turn in the spoons as well as the scepter and crown only when..

5.7 Lexicographic information

The template further records standard lexicographic information.

- Semantic field/domain
  The lexicographers record a semantic domain label from a list of labels, such as "food," "mental state," and "work."

- Semantic (paradigmatic) relations to other idioms.
  Here we record idioms that are synonyms, hypernyms, hyponyms, antonyms, etc. of the entry. Some semantic domains (e.g., death, mental states) are richly lexicalized and have paradigmatically related idioms.

- Register/Style
  We categorize the usage of the idiom as "ironic," "disparaging," "humorous," etc.

- Diachronic information
  The 100-million word corpus covers the entire 20th century. Changes in the form and meaning of idioms during this time window can be observed and are recorded in the template.

Finally, the linguists and lexicographers can add open format comments in a box provided for that purpose.

6 Conclusion and outlook

Currently, over 500 verb phrase idioms have been analyzed in depth. The data cited in this paper give an indication of the range of variation shown by idioms, raising the question as to the lexical and grammatical status of idioms. They also show that idioms cannot be treated lexicographically in

the same way as ordinary single-word lexemes or truly frozen sequences like *by and large*.

We have described the methods for constructing a new kind of lexicographic resource, which combines lexico-linguistic analyses with relevant corpus data. The feedback loop — corpus data enables analysis, and analysis is linked to corpus data — ensures an empirically based transparent analysis. We hope that our model will inspire similar resources in other languages or focusing on different parts of the lexicon. We anticipate making the data available for further research at the end of the project.

7 Acknowledgment

This work was supported by the Wolfgang Paul-Preis of the Alexander-von-Humboldt Foundation's Zukunftsinvestitionsprogramm awarded to Fellbaum. The following project members have greatly contributed to the work described here: Undine Kramer, Gerald Neumann, Ekatherina Stathi, Alexej Sokirko, Anna Firenze, Elke Gehweiler, Axel Herold, Renata Kwasniak, Christoph Srubar, Anne Urbschat, Kai Zimmer.

References

[1] Dobrovolskij, Dmitri (1999). "Haben transformationelle Defekte der idiomstruktur semantische Ursachen?" In: Fernandez-Bravo, Nicole, Behr, Irmtraud, and Rozier, Claire (Eds.): Phraseme und typisierte Rede. Uerogermanistik Vol. 14, Tübingen, Stauffenburg, 25-27.

[2] Fellbaum, C. (2004). "Idiome in einem Digitalen Lexikalischen System." Linguistik und Literatur, Vol. 34, no. 136, 56-71.

[3] Fellbaum, C., Kramer, U., and Stantcheva, D. (2003). "Etwas and Eins in VP-Idiomen." Proceedings of Annual Meeting of the Institute for German Language, Mannheim, Germany.

[4] Fellbaum, C. (2002b). "VP Idioms in the Lexicon: Topics for Research Using a Very Large Corpus." In: Busemann, S., (Ed.), Proceedings of KONVENS 2002. Saarbrücken, Germany: DFKI.

[5] Fellbaum, C., Kramer, U., and Neumann. G. (2005). "Corpusbasierte lexikographische Erfassung und linguistische Analyse deutscher Idiome." In: Proceedings of the Euraphras Meeting, Basel, Switzerland, 183-199.

[6] Fellbaum, C., and Stathi. E. (2006). "Ididome in der Grammatik und im Kontext: We brüllt hier die Leviten?" In: Proost, K., and Winkler, E. (Eds.) Festschrift für Gisela Harras. Berlin: de Gruyter.

[7] Fellbaum, C. (to appear). "The Ontological Loneliness of Idioms." In: Schalley, A., and Zäfferer, D. (Eds.) Ontolinguistics, Mouton.

[8] Fleischer, W. (1997). Phraseologie der deutschen Gegenwartssprache. 2nd edition, Tübingen: Niemeyer.

[9] Geyken, A. and Klein, W. (2004). Projekt "Digitales Wörterbuch der deutschen Sprache des 20. Jh.". In: Jahrbuch der BBAW 2003, Berlin: Akademie Verlag.

[10] Geyken, A. (2004). Korpora als Korrektiv für einsprachige Wörterbücher. In: Zeitschrift für Literaturwissenschaft und Linguistik, Jg. (2004), H. 136, S. 72-100.

[11] Geyken, A., Sokirko, A., Rehbein, I., and Fellbaum, C. (2004). "What is the optimal corpus size for the study of idioms?" Annual meeting of the German Linguistic Society, Mainz 25.-27.02.2004.

[12] Glucksberg, S. (1993). Idiom Meaning and Allusional Content. In: Cacciari, Cristina, and Tabossi, Patrizia: The comprehension idioms. Hillsdale, NJ: Erlbaum.

[13] Herold, A. (2005a). Suchanfragen: Dokumentation zur Korpusabfrage und zur Arbeit mit der Idiomdatenbank. Internal ms., Berlin-Brandenburg Academy of Sciences.

[14] Herold, A. (2005b). Reducing the Size of Sample Corpora for Research on Idioms in the German Language. Poster presented at the Conference on Corpus Linguistics, University of Birmingham, UK, July 2005.

[15] Kramer, U., Neumann, G., Stathi, K. and Fellbaum. C. (2005). "Kollokationen im Wörterbuch"–Das Wolfgang Paul-Preis Projekt an der Berlin-Brandenburgischen Akademie der Wissenschaften. *Zeitschrift für Germanistik*.

[16] Lezius, W. (2000). Morphy - German Morphology, Part-of-Speech Tagging and Applications. In: Ulrich Heid; Stefan Evert; Egbert Lehmann and Christian Rohrer, editors, Proceedings of the 9th EURALEX International Congress pp. 619-623 Stuttgart, Germany.

[17] Moon, R. (1998). Frequencies and Forms of Phrasal Lexemes in English. In: Cowie, A. (Ed.) Phraseology: Theory, Analysis, Applications. Oxford: Oxford University Press.

[18] Neumann, G., Fellbaum, C., Geyken, A., Herold, A., Hümmer, C., Körner, F., Kramer, U., Krell, K., Sokirko, A., Stantcheva, D., and Stathi, K. (2004). A Corpus-Based Lexical Resource of German Idioms. In: Proceedings of the Workshop on Electronic Lexicons, eds. P. Saint Dizier and M. Zock, COLING, Geneva, 48-52.

[19] Nunberg, J., Sag, I., and and Wasow, T. (1994). Idioms. Language 70:491-538.

[20] Schemann, H. (1993). Deutsche Idiomatik. Die deutschen Redewendungen im Kontext. Stuttgart and Dresden.

[21] Sokirko, A. (2003). DDC - A Search Engine For Linguistically Annotated Corpora. In Proceedings of Dialogue 2003 (Protvino, Russia, June 2003).

[22] Sokirko A. (2004). Morphological components on www.aot.ru. In "Proceedings of Dialogue 2004" (Russia, Verchnevolzhskiy)