Project Note

A Corpus-based Lexical resource of German Idioms

G. Neumann, C. Fellbaum, A. Geyken, A. Herold, C. Hümmer, F. Körner, U. Kramer, K. Krell, A. Sokirko, D. Stantcheva, E. Stathi

Project: "Collocations in the German Language",
Berlin-Brandenburg Academy of Sciences and Humanities
Jägerstraße 22/23
Germany, 10117 Berlin
gneumann@bbaw.de

Abstract

In this paper, we present the design of a lexical resource focusing on German verb phrase idioms. The entry for a given idiom combines appropriate corpus examples with rich linguistic and lexicographic annotations, giving the user access to linguistic information coupled with the appropriate attested data and laying the groundwork for empirical, reproducible research.

1 Introduction

The project "Collocations in the German Language", located at the Berlin-Brandenburg Academy of Sciences, studies the properties of verb phrase idioms such as *jmdm. einen Bären aufbinden* (lit. 'tie a bear onto sb.'s back,' i.e. 'tell a big lie') and *einen Bock schiessen* (lit. 'shoot a buck,' i.e. 'commit a grave error').

Idioms are sometimes referred to informally as "long words" and are treated as fixed strings with no internal structure. While some idioms (like the much-discussed English example *kick the bucket*) have semantically opaque components and do not undergo syntactic or lexical alternations, others (like *bury the hatchet*) behave more like freely composed VP.

Nunberg, Sag, and Wasow (1994) and Dobrovolskij (1999) are among the linguists who have observed a correlation between an idiom's fixedness and its semantic opacity: The more fixed an idiom, the more semantically opaque it can be. But this correlation is not always straightforward, and to understand the way speakers represent idioms in their internal grammars we must investigate the full range of behavior. For example, the German equivalent of *kick the bucket* is lit. 'bite into the grass' (*ins Gras beissen*). While *Gras* (just like 'bucket') does not seem to be mappable onto the concept, it can be modified with an adjective, as our corpus example 'bit the Texan grass' shows. In fact, the adjective must be interpreted as having scope over the entire VP, not just the NP, as the phrase refers to someone dying in Texas.

Lexical substitution and variability of the idiom's components in general can show that an element is assigned a particular interpretation, as in the (diachronically differentiated) variants *sich auf die Strümpfe machen* and *sich auf die Socken machen* (lit. 'make oneself on the stockings/socks', i.e. 'to get going or get moving'). Here, the substitution of a near-synonym indicates that speakers assign some meaning to the noun, most likely the footwear that is associated with travel on foot. But speakers make substitutions without necessarily assigning a meaning to the constituents of the idiom.

Current lexicographic and linguistic treatment of VP idioms does not attempt to reflect the full range of the idioms' properties or classify them accordingly. Our goal is to give a data-oriented, comprehensive linguistic account of a broad spectrum of German idioms.

The empirical basis for our work is the corpus of the Digitales Wörterbuch der deutschen Sprache (DWDS), a corpus of almost 1 billion words from texts drawn from a variety of genres and covering the entire 20th century.

We are creating a resource that combines features of a dictionary, a grammar of idioms, and a corpus. Central to our methodology are two main components: an electronic knowledge base created via structured annotations, and a corresponding sub-corpus of examples (example corpora) for each entry.

2 The Example Corpora

Our lexicon combines subtle linguistic annotations with specific corpus data.

For each target idiom, a sub-corpus, containing appropriate examples drawn from the 1-billion-word-corpus, is created. Following the identification of a target idiom, corpus queries are written to generate the candidate set of relevant corpus data, taking advantage of a search engine developed in-house. The queries include specific lexical and morphological elements as well as Boolean operators. The resultant candidate example corpus, in XML/TEI format, is manually inspected for false positives, i.e. strings which are not instances of the idiom but rather accidental products of the query.

The number of true positives gives an idea of the frequency of idioms in the language – information that was not easy to gather until now. For example, the idiom *jmdn. zur Minna machen* (lit. 'make sb. a Minna,', i.e. 'to publicly reprimand sb.') is attested only 41 times, while *ein Auge auf etw. haben* (lit. 'to keep an eye on sth.') yields 560 examples.

To ensure reproducibility of result sets, examples are marked with tags identifying their corpus query and the particular corpus version. Thus we have the possibility of comparison between the manually inspected example corpora and the dynamically increasing DWDS-Corpus.

3 The Annotation Template

We created a template for linguistic and lexicographic annotation, a kind of digital questionnaire. It serves as both input and output interface and links the annotated example corpora with the idiom knowledge base. The data entry interface supports a structured entry created by the linguists/lexicographers, who record various properties of the idioms as evidenced by the examples.

The template design reflects the information which is considered relevant in the discussion of idioms from a lexicographic and linguistic point of view, focusing on the interplay between normal usage and flexibility. The template consists of the following major parts:

The first part contains information which is typically found in dictionaries. This includes the citation form of the idiom, together with a definition. A difference from standard dictionaries is the link to the data. We record the first and last occurrence in the corpus and in this way we indicate the time span of corpus examples of the idiom during the 20th century. The citation form is derived on the basis of the corpus evidence according to statistical significance (cf. example below). The typical usages lead to the formulation of the citation form, which is linked to some of them. In addition, there is information on alternate forms (e.g. different aspectual varieties of an idiom, transitive vs. intransitive uses, etc.) with a reference to the corresponding entry.

The second part is the syntactic structure of the idiom, in particular the structure of the VP including the subject. We show dependencies between constituents on three levels (two phrase levels and a terminal level), using the category variables of a German tag set[1]. At the terminal level we fill in the lexical material which corresponds to the citation form. Finally, we note the status of each component in the idiom according to degree of fixedness (external argument, core component, obligatory, optional, etc.).

This 'tree' structure automatically creates a table which records the morphosyntactic properties of idiom components. Since idioms are considered as typically restricted according to morphosyntactic properties (e.g., noun complements may occur only in the singular or only in the plural), we explore to which extent this is supported by the data.

On the basis of these morphosyntactic regularities, we derive the citation form. This is considered to be the normal (or typical) form, according to frequency of usage. We are also interested in non-typical or "deviant" usages. The term should be understood not in the sense of "abnormal" or "ungrammatical", but as infrequent or not statistically significant (sometimes idiosyncratic, sometimes more widespread), yet important for insight into the linguistic properties of an idiom.

Idioms are also subject to restrictions concerning the passivization of the nominal components, their pronominalization, relativization, and other syntactic processes sometimes called "transformations". If we take the example mentioned earlier, one would not expect to encounter **das Gras, in das er gebissen hat* (relativization), **das Gras wurde gebissen* (passivization), etc. In reality, these transformations can be observed in most cases. We record them in a separate table and link every transformation type to the corresponding occurrences in the example corpora.

The template also offers the possibility to include semantic as well as historical information, if diachronic changes or noteworthy developments can be observed in the corpus data.

By way of illustration, let us consider an example, focusing attention on variability. In

[1] STTS (Stuttgart-Tübingen Tagset)

Modern German the noun *Bockshorn*[2] is considered to be a phraseologically bound word, i.e. it does not occur in isolation, but only in a fixed expression[3]. Idioms with phraseologically bound words are said to display the highest degree of fixedness and absence of flexibility. We want to test this hypothesis against corpus data.

The phraseological dictionary "Duden (Vol. 11)" gives s.v. *Bockshorn* the citation form *jmd. lässt sich nicht ins Bockshorn jagen*, which is roughly equivalent to English 'refuse to allow oneself to be intimidated'. The "Wörterbuch der deutschen Gegenwartssprache" (WDG) records also the transitive alternation *jmdn. ins Bockshorn jagen*, which might be translated as 'to put the wind up sb.'

The following table shows the results which were given by three different queries to the corpus for this idiom:

Query	*Number of hits in the corpus*
Bockshorn	285
*horn && jagen && !@Bockshorn	27
((Bockshorn* && !@Bockshorn) || *bockshorn)	35
Total	347

Table 1: Queries and hits for *Bockshorn* in the DWDS-Corpus

As the table shows, the queries leave open the possibility of finding variation at every slot: The position of the verb is left unspecified in the first query. In the second query we try to find instances where the first component of the noun (*Bock-*) might be substituted by another noun[4], while the third query should deliver instances of compounds with *Bockshorn* (cf. the form *Bockshornklee* in Table 2 below).

Linguistic analysis of the data led to the formulation of two alternate citation forms and consequently two entries for the idiom:

1) jmd. lässt sich nicht ins Bockshorn jagen (192 hits)
2) jmd. jagt jmdn. ins Bockshorn (44 hits).[5]

The first is a form with the so-called "*lassen*-passive". It is much more frequent than the transitive form; actually it is the typical form associated with *Bockshorn*. The results show a remarkable uniformity of the idiom components, as the following table shows:

Total		***192***
Noun	**Bockshorn**	**182**
	Boxhorn	8
	Bockshornklee	1
	Hasenhorn	1
Preposition	**ins**	**191**
	in[6]	1
Infinitive	**jagen**	**191**
	jache[7]	1
Verb	lassen	192
Negation	**Overt negation** (different possibilities)	**141**
	Irrealis	15
	Affirmation	36

Table 2: Distribution of components for the "*lassen*-passive" alternant

The forms which are mentioned first in Table 2 (bold face) are seen as the "normal" components in the sense of statistical significance, the other forms are considered as "deviations" from the norm. These include the substitution of components. Cf. the noun: The form *Boxhorn* is phonologically identically to *Bockshorn*; perhaps its usage represents the need to motivate the original noun.[8] The substitution of *Bockshornklee* can be explained with reference to the context: This noun appears earlier in the text and triggers the use of the idiom, a not unusual phenomenon in journalese texts. Finally, *Hasenhorn* shows the typical pattern of noun substitution in idioms in German, i.e. the substitution of one part of a compound noun typically by a synonym or a word belonging to the same lexical field (*Hase-* 'rabbit' for *Bock-* 'buck'). Again, this substitution is triggered by the topic of the text. Note that this substitution occurs

[2] The etymology of the word is obscure. Röhrich (2001) offers nine etymologies for the idiom according to different possible sources of the noun. He reaches no conclusion. Synchronically, the word is identical to the name of a low-growing plant (fenugreek) with tough wiry stems, also called *Bockshornklee*. But it can also literally denote the horn of a goat (*Bock* 'male goat' and *Horn* 'horn').

[3] Moon (1998) terms these 'cranberry' collocations. Actually, this particular word (*Bockshorn*) occurs outside a fixed expression, but it occurs only in highly technical contexts (cf. footnote 5 below).

[4] The opposite case, i.e. looking for cases with the component stable (*-horn*), gave no results.

[5] 111 hits were non-idiomatic uses of the noun, especially in popular medicine and compounds.

[6] Typing error

[7] Dialectal form for *jagen*.

[8] An alternative interpretation relates the form *Boxhorn* to 15th century words relating to God (cf. Röhrich 2001).

even though the speaker cannot assign a meaning to the lexeme *Bockshorn* (cf. Introduction).

Now consider the figures for the transitive alternant *jmd. jagt jmdn. ins Bockshorn*:

Total		***44***
Noun	**Bockshorn**	**41**
	Boxhorn	1
	Bockhorn	1
	Gruselhorn	1
Preposition	**ins**	**43**
	in's	1
Verb	**jagen**	**42**
	einjagen	1
	führen	1

Table 3: Distribution of components for the transitive alternant

The citation form given is derived on the basis of these quantities, which distinguish normal usage from "deviant" usages. Beside the form *Boxhorn* (cf. above) *Bockhorn* occurs once.[9] *Gruselhorn* (*Grusel-* 'scary') is a nonce creation for the expression of emphasis since it combines the meaning of the idiom and the meaning of 'to scare sb.'

As for the verb, it is substituted twice. The verb *führen* 'to lead' fulfils the same semantic function as *jagen* in this idiom, i.e. the expression of causativity. The difference is that *führen* is unmarked: In German it is often used as a function verb to express causativity (to cause a change of state for sb.). The substitution of *einjagen* on the other hand may be due to its occurrence in the near-synonymous expression *(jmdm.) Angst/Schrecken einjagen* 'to scare sb.'

To sum up, if variability correlates with flexibility (or fixedness) of an idiom, then these figures lead to the conclusion that idioms with phraseologically bound forms are typically fixed. But contrary to claims in the literature that there is no variability at all, corpus evidence shows that there is some patterned variability. The interesting question is whether it is also predictable.

Finally, it should be noted that the postulation of two entries (the *lassen*-Passive and the transitive/ causative) are necessary for two reasons. Firstly, the two entries differ in syntactic structure (argument structure, negation, etc.). This difference automatically forces us to establish separate entries, because the whole design of the template is based on syntactic structure. This means that the subsequent sections (morpho-syntactic properties and variations or deviations from the citation form) are created automatically on the basis of the syntactic structure of the idiom. Of course, they have to be filled in manually, but their design is uniform for all idioms which ensures comparability.

Secondly, the *lassen*-Passive (plus negation and reflexivity) as a typical instance of this idiom is not a regular (or necessary) correlate of the transitive form, as a pure passive form would be. Moreover, the two entries differ pragmatically.

[9] It could be an error or just omission of the connecting *-s-*.

4 The Knowledge Base

The information recorded in the template is stored in a MySQL database, which constitutes the actual knowledge base for German idioms. This knowledge base can be queried in various different ways. In particular, it is possible to query for syntactical structures of idioms and for all kinds of variation from the citation form found in the example corpus.

We developed a label language that permits a precise, automatic assignment of examples to most of the properties recorded in the template. Thus the lexicographers' decisions preserve a high degree of transparency.

5 Use of Resource

The contents of the knowledge base will be made available via the Internet. The template, in its function as a user interface, is browser-based and directly accesses the database.

Users can search for specific idioms (or substrings thereof), examine their linguistic properties, and find the appropriate corpus examples.

This automatic linking of examples to the corpora makes the analysis transparent in a way that is not possible in conventional dictionaries.

6 Conclusion

In sum, our resource combines properties of a dictionary and a grammar with a corpus. A learner or general user can look up a specific idiom and study its properties. A linguist can search for specific linguistic structures (passives, clefts, etc.) and find those idioms that show the features in question. Other users will no doubt find other modes of application.

7 Acknowledgements

This work is supported by the Alexander von Humboldt Foundation's Zukunftsinvestitions-programm through the Wolfgang Paul-Preis.

References

Dobrovol'skij, Dmitrij. 1999. "Haben transformationelle Defekte der Idiomstruktur semantische Ursachen?" In: Fernandez-Bravo, Nicole, Irmtraud Behr und Claire Rozier (eds.) 1999: *Phraseme und typisierte Rede* (=Eurogermanistik 14). Tübingen: Stauffenburg, 25-37.

Dobrovol'skij, Dmitrij. 2002. "Zum syntaktischen Verhalten deutscher Idiome (am Beispiel der Passivtransformation)" In: Wiesinger, Peter (ed.) 2002: *Akten des X. Internationalen Germanistenkongresses Wien 2000. "Zeitenwende - Die Germanistik auf dem Weg vom 20. ins 21. Jahrhundert". Jahrbuch für Internationale Germanistik. Reihe A, Kongressberichte Bd. 54, Band 2, Entwicklungstendenzen der deutschen Gegenwartssprache; Lexikologie und Lexikographie*. Bern et al.: Peter Lang, 379-384.

Duden, Redewendungen: Wörterbuch der deutschen Idiomatik. 2002. Hrsg. von der Dudenredaktion. [Red. Bearb.: Brigitte Alsleben; Werner Scholze-Stubenrecht]. - 2., neu bearb. und aktualisierte Aufl.. Mannheim et al.: Dudenverlag.

Fellbaum, Christiane, Undine Kramer und Diana Stantcheva. 2004. "*Eins, einen, eine* und *etwas* in deutschen VP-Idiomen" In: Steyer, Kathrin (ed.) 2004: *Wortverbindungen - mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003*. Berlin, New York: Walter de Gruyter, 167-193.

Fleischer, Wolfgang. 1997. *Phraseologie der deutschen Gegenwartssprache*. 2., durchgesehene und ergänzte Auflage. Tübingen: Max Niemeyer.

Moon, Rosamund. 1998. "Frequencies and Forms of Phrasal Lexemes in English". In: Cowie, A. P. (ed.) 1998. *Phraseology: Theory, Analysis, Applications*. Oxford: Oxford University Press, 79-100.

Nunberg, Geoffrey; Ivan A. Sag und Thomas Wasow. 1994. "Idioms" *Language 70/3*, 491-538.

Röhrich, Lutz. 2001. *Das große Lexikon der sprichwörtlichen Redensarten* [Elektronische Ressource]. Darmstadt: Wissenschaftliche Buchgesellschaft.

Wörterbuch der deutschen Gegenwartssprache. 1967-1977. Akademie der Wissenschaften der DDR, Zentralinstitut für Sprachwissenschaft. Hrsg. von Ruth Klappenbach und Wolfgang Steinitz. Berlin: Akademie-Verlag.

Relevant web pages:
http://www.dwds.de
http://www.bbaw.de/forschung/kollokationen/