Computer-aided Analysis of Idiom Modifications in German

Elena Krotova
Institute of Linguistics, Russian Academy of Sciences
E-mail: elena_krotova@inbox.ru

Abstract

This paper deals with corpus approaches to the study of modifications of idiomatic expressions in German. It concentrates on one group of phraseological units, idioms. In spite of a high degree of stability, idioms still undergo different modifications. To get reliable results about idiom modifications, a large number of modified target structures is crucial. Therefore, a Python-program has been created that obtains information about the usage of idioms and about their possible modifications from a corpus. It also summarizes the data in the form of graphs. The report will look further into the program's opportunities to acquire information about idiom usage, how idiom modifications correspond to the syntactic behavior of their paraphrases or free phrases containing the same verb as the idiom under discussion and in what ways such data can facilitate the work of a phraseologist.

Keywords: corpus linguistics, phraseography, modifications of idiomatic expressions

1 Introduction

The compilation of dictionary entries for phraseological units, especially for idiomatic expressions, is difficult for many reasons.[1] Idioms have a complex semantic structure and a number of syntactic peculiarities that affect their usage in speech. Only with a detailed description of idioms' semantics and syntax can non-native speakers learn to use phraseological units in an appropriate, native-like way. Before large machine-readable corpora appeared, makers of phraseological dictionaries had to rely largely on their own language intuition, whereas now it is possible to verify information using corpora. For frequent idioms we can find thousands of occurrences in large corpora. This amount of material is sufficient to describe in detail the behavior of idioms in written language. The problem however is that the analysis of such an amount of data takes a lot of time, especially if it is not a detailed study of one particular idiom, but the compilation of hundreds of articles for a phraseological dictionary. This time could be reduced through the use of automatic methods that analyze corpora data, search for different idiom modifications and summarize the obtained information. In the following, we will discuss the development of such a program for frequent idioms of the German language and present its first results.

2 Idiom Modifications

Idioms possess a high degree of stability, but they still undergo different modifications. In Figure 1 you can see two graphs[2] with modifications of German idioms:

1 I use the term idiomatic expressions, or idioms, to refer to "phrasemes with a high degree of idiomaticity and stability. In other words, idioms must be fixed in their lexical structure (however, this does not exclude a certain variation), and they must be, at the same time, semantically reinterpreted units (i.e. they do not point to the target concept directly but via a source concept) and/or semantically opaque" (Dobrovol'skij 2006; for detailed explanation of the term idiomaticity see Baranov & Dobrovol'skij 2008: 50).

2 I use the term idiomatic expressions, or idioms, to refer to "phrasemes with a high degree of idiomaticity and stability. In other words, idioms must be fixed in their lexical structure (however, this does not exclude a certain variation), and they must be, at the same time, semantically reinterpreted units (i.e. they do not point to the target concept directly but via a source concept) and/or semantically opaque" (Dobrovol'skij 2006; for detailed explanation of the term *idiomaticity* see Baranov & Dobrovol'skij 2008: 50).

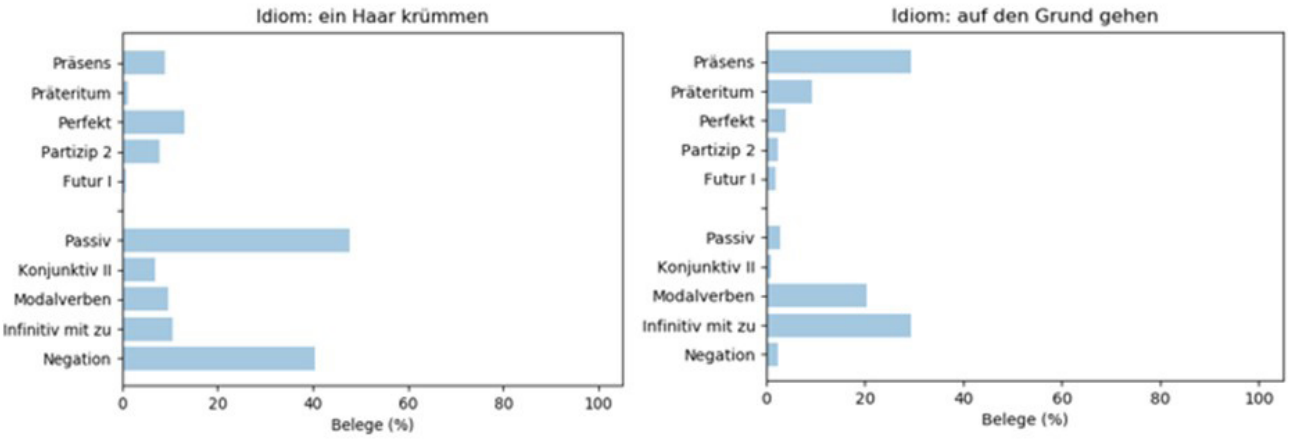

Figure 1: Idioms *ein Haar krümmen jmdm., auf den Grund gehen etw.*

The first idiom *ein Haar krümmen jmdm.* 'to lay a finger on sb.' is mostly used in the passive voice and with negation, while for the second idiom *auf den Grund gehen* 'to drill down on sth.' these modifications are not frequent. The second idiom is mostly used in infinitive constructions, in present and with modal verbs.

We distinguish between the following kinds of modifications: morphological, lexical, lexical-syntactic and syntactic (see Dobrovol'skij 2008: 308-309). Here are some examples from German:

- Morphological modifications include article variation (*auf dem [einem] absteigenden Ast sein* 'to be on the downgrade').
- Lexical modifications include changes in component structure (omission or addition of an element, its substitution by another word). Example: *von allen guten Geistern verlassen sein* 'to have taken leave of one's senses' and its lexical variants *von guten Geistern verlassen sein, von allen Geistern verlassen sein,* seldom *wie von allen guten Geistern verlassen sein,* seldom *von allen guten Göttern verlassen sein*.

Lexical modifications are complicated to detect, because a researcher does not always know about all the lexical changes an idiom can undergo[3].

- The lexical-syntactic type covers modifications which affect both syntactic and lexical idiom structure (e.g. putting an adjective modifier between an article and a noun). It can refer to the adverbial (e.g. *sein Herz ausschütten jmdm.* 'to open one's heart' modified by *so richtig* 'really': *so richtig das Herz auszuschütten*), the attributive (e.g. *einen Haken haben: (etw.) hat einen Haken* 'There is a hitch somewhere' modified by *klein* 'small': *einen kleinen Haken hab*en) and to the so-called metalinguistic type (e.g. *sich fühlen wie ein [der] Fisch im Wasser* 'to be in one's element' modified by *viel zitiert* 'much-cited': *wie der viel zitierte Fisch*).
- Syntactic modifications contain the use of idioms with negation, in passive voice, questions, infinitive constructions and so on.

This study deals mostly with syntactic modifications as well as tenses the verbal component of the idiom is used in. The extracted data is presented on the graphs. Other types of modifications are analyzed only to some extent (see Section 4).

There are idioms that hardly ever show modification, such as *abwarten und Tee trinken* 'to wait and see':

3 See Section 4 for further elaboration on the graphs.

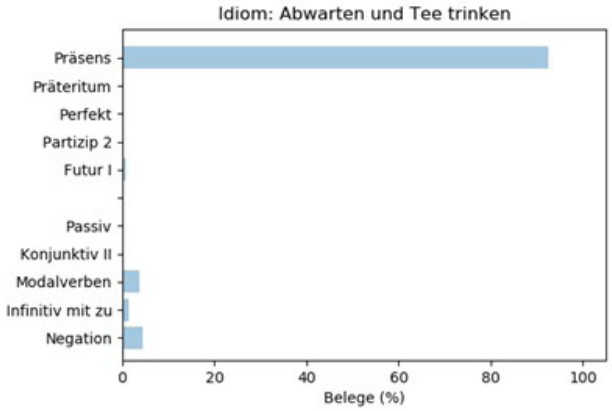

Figure 2: Idiom *Abwarten und Tee trinken*

However, it is not possible to say that this idiom has an absolutely fixed form. There are occurrences where it is used with negation, with modal verbs and in infinitive constructions. The verbal component of the idiom *trinken* is used in 91.3% of texts in this particular word form, in 3% with a modal verb and in 2.4% in infinitive constructions with *zu*.

This case is rather rare. Among the 100 idioms analyzed, such a small number of changes has been found only for the idiom *dem Fass den Boden ausschlagen: Das schlägt dem Fass den Boden aus* 'That's outrageous'. This idiom, unlike the previous one, can be used in other tenses, although such occurrences are rare, and the verb can change its form in the present (though the word forms *schlägt aus* and *ausschlägt* comprise up to 86% of occurrences). To the right of the idiom there is a graph for a free phrase containing the verb that is a part of the idiom:

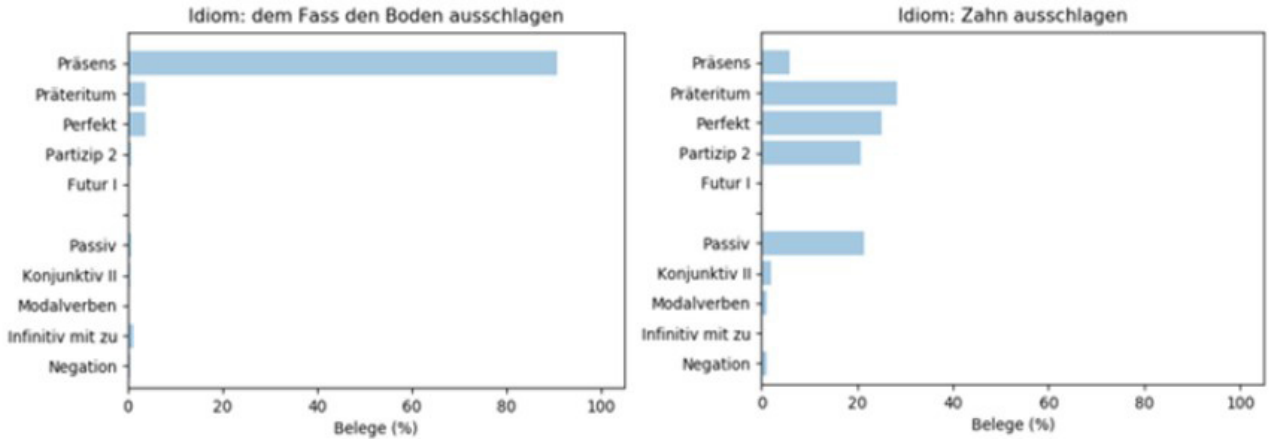

Figure 3: Comparison of the idiom *dem Fass den Boden ausschlagen* and the free phrase *jmdm. einen Zahn ausschlagen*

As you can see, although the idiom is seldom used in the passive, its verbal component in other free phrases is used in the passive quite frequently. The structure of the idiom on its own does not restrict such a modification, e.g. the sentence *Dem Fass wurde der Boden ausgeschlagen* would be correct from a purely grammatical point of view. However, such a modification of the idiom hardly ever occurs in speech. Moreover the free phrase *den Zahn ausschlagen* seldom occurs in the present, which can be connected with pragmatics: in the prototypical situation the phrase describes a result of a physical action, so the usage in past tenses is more probable.

Therefore, a lexicographer should provide language learners with information about the modifications an idiom undergoes, because this information cannot be inferred from the syntactic behavior of the idiom's components in free phrases.

3 Material

To get reliable results about idiom modifications a researcher needs to acquire and analyze as many text examples as possible. It is clear that the biggest corpora are collections of texts from the Internet that can be searched by a search engine. However, this approach has its limitations, because a search engine is focused on information retrieval and not on the extraction of linguistic information. That is why linguistically annotated text corpora have been chosen for the study of idiom modifications. The biggest corpus of the German language is *Deutsches Referenzkorpus* (DeReKo), which contains more than 42 billion tokens (03.02.2018), but it is unbalanced and consists mainly of newspaper texts. This has obvious disadvantages: 1) not all idioms are used in written speech and particularly in newspaper texts, even if they are frequent in spoken language, 2) we often have wordplay with idioms in newspapers that does not always reflect their usage in speech. However, the fact that the corpus is large and we can find hundreds of each phraseme's occurrences in modern texts outweighs this drawback.

In order to have more homogenous text material a virtual corpus has been created for research purposes. The corpus contains only newspaper articles published in Germany after 1980 and excludes texts published in other German-speaking countries for the following reasons: Austrian and Swiss texts can contain some idiom modifications which can be typical only for the particular national variant of German but not for other variants; since the modern usage of idioms is investigated, only texts published since 1980 are considered; the corpora contain very few fiction texts published since 1980 (less than 0.01% of the whole text archive), so they are excluded to make it clear that the analysis is based only on newspaper texts.

At the starting stage of the project, 100 widely used German idioms were chosen. Most of them have the structure verb + preposition + noun. The surveyed phraseological units come from the dictionary of frequent German idioms (Dobrovolskij 1997); their frequency was checked against DeReKo. The information given in the dictionary about possible lexical, morphological and lexical-syntactic modifications of the idioms was obtained through the analysis of different lexicographic resources and text corpora.

4 Methodology

DeReKo queries have been formulated manually for all of the selected idioms. The queries are made in such a way that possible modifications are not excluded from the results[4]. E.g. the idiom *auf den Grund gehen* 'to get to the bottom of sth.' During the work on the dictionary of frequent German idioms it turned out that the idiom under discussion can be used in the passive form, but it does not seem

4 Lexical modifications can be explored in different ways. Besides consulting different lexicographic resources a researcher can omit some components of an idiom and try to find out what other words can fill the gap. In DeReKo one can use a tool for co-occurrence analysis (*Kookkurenzanalyse*). E.g. when we search for co-occurrences for the phrase *auf den Grund*, we see that the most statistically significant co-occurrence partner will be *gehen*. But if we search for co-occurrences for the verb *gehen*, we won't find the phrase *auf den Grund*, may be because of the very high frequency of the verb. Another example for the idiom *von allen guten Geistern verlassen sein*: if we search for *von allen guten Geistern* the most significant co-occurrence partner will be *verlassen*. If we search for *von allen guten* and *verlassen*, we will find the lexical modification *von allen guten Fußballgeistern verlassen sein*. It is not frequent (only ten occurrences from more than 1,800 sentences containing the idiom), but statistically significant, because this word occurs very seldom, yet in about 30% of sentences together with *verlassen*.

to have any frequent morphological or lexical modifications. As a result, the query looks as follows: *(auf /+w3,s0 Grund) / s0 &gehen*. It means that *gehen* can be used in any form (operator &) and have any position in a sentence (operator /s0 means that the verb must be in the same sentence as the prepositional group). Besides, a maximum of two tokens can appear between *auf* and *Grund* (operator *w*), *auf* is followed by *Grund* (operator +), and both tokens must occur within the same sentence (operator /s0). The paragraphs where the idioms occur can be exported. DeReKo doesn't allow you to export over 10,000 examples, but this number is enough for studying the usage of idioms in written texts.

The queries have not been automated for the following reasons: in every case the researcher should decide how to formulate queries so that the data obtained contain a minimum amount of sentences where the idiom's components are used literally and not idiomatically, but at the same time do not exclude possible modifications. For instance, the idiom *sich (D) ins Fäustchen lachen* 'to laugh in one's sleeve' contains the noun *Fäustchen*, which is not that frequent. That is why it will be enough if we just search for sentences where both components occur together. The situation is different for the idiom *mit den Wölfen heulen* 'to do in Rome as the Romans do'. If we search for the lemmas[5] *Wolf* 'wolf' and *heulen* 'to howl' occurring in the same sentence, we will find a lot of targets where there is no idiom. Therefore, the query should be restricted. For example, *(mit /+w3,s0 Wölfen) /s0 &heulen*, which means that the lemma *heulen* occurs in one sentence with word forms *mit* and *Wölfen*, the maximal distance between them makes two words.

The specially designed program is used to analyze the extracted data. The obtained data is provided at bitbucket.org (Deutsche Idiomatik). It targets the following information:

- tenses and word forms the verb of the idiom is used in (*Präsens, Präteritum, Perfekt, Futur I*[6])
- syntactic modifications the idiom undergoes, such as usage in passive voice; in *Konjunktiv II*; accompanied by modal verbs and in infinitive constructions. Verbs (*werden* in passive voice, verbs in conjunctive mood, modal verbs) can be used in any tense. Such contexts do not overlap with the first group. For instance, if there is an idiom in the passive voice present tense, it will only be counted as an idiom in passive voice. The program also searches the contexts with negation *nicht* or *kein*.

Here is an example: 805 contexts have been found for the idiom *jmdm. ein Armutszeugnis ausstellen* 'to show sb.'s incapacity / incompetence / shortcomings', among them 56% in *Präsens*, 14.66% in *Perfekt*, 6.58% with *Partizip II* (auxiliary verbs *sein* or *haben* have not been found), 1.61% in passive voice, 4.98% in *Konjunktiv II*, 4.59% with modal verbs, 1.5% in infinitive constructions. This totals 99.72%. In addition, future tense *Futur I* takes up about 0.3%.

The most frequent verb form is *stellt aus* (28.9%). The idiom seldom occurs with negation (4.59%). The most frequent modal verb is *können* (64% of all occurrences with modal verbs).

Moreover, the tokens preceding the nouns in the nominal component are analyzed. E.g. the noun *Armutszeugnis* is preceded by the article *ein* in 75% of all examples. Other variants are as follows: definite article *das* (1.49%), lexical and syntactic modifications: such words as *solches*, *dieses* (about 0.8% each) and such adjectives as *politisches* (1.4%), *größeres*, *großes* (0.8% each), *geistiges*, *eigenes* (0.6% each). By these means the researcher acquires information about possible morphological (*ein* or *das*), lexical and syntactic modifications (*solches, dieses politisches*, *größeres*, *großes*).

The program application can go further to search for sentences with questions and those where an idiom is used in the first person and with metalinguistic constructions (see metalinguistic type in Section 2).

5 In order not to exclude syntactical or morphological modifications the queries normally don't contain articles before the nominal component and contain all possible verb forms. Information about lexical modifications was obtained through the previous research. If an idiom occurs in several lexical variants, separate searches are made.

6 Lemma is understood as a set of words.

In the end, the program creates files that contain an idiom's occurrences in different tenses and with the modifications mentioned above. For all modifications there are separate text documents, so the researcher can analyze each of them in detail. The program also summarizes the data in the form of graphs.

5 Results

A total of 100 German idioms have been analyzed. The number of idioms is not very high now, but still the attempt has been made to find idioms with similar modification profiles. The following results have been obtained:

- Almost a third of idioms (related lists and graphs can be found in the project folder) are used in more than 15% of occurrences with modal verbs.
- 10% of idioms are used in infinitive constructions in more than 20% of texts.
- 5% of idioms are used with negation in over 25% of examples.
- 15% of idioms appear in present tense in more than 60% of occurrences.
- 6% of idioms are used in *Präteritum* more often than in *Präsens* and *Perfekt*.

The groups were established manually because the number of analyzed idioms is rather small at the moment, and as this number increases in the future appropriate statistics will be obtained.

Let us now take a closer look at the idioms used mostly in *Präsens*. In addition to the idiom mentioned in the introduction *abwarten und Tee trinken* 'just wait and see', this group includes several idioms with similar semantics ('to get on sb.'s nerves'):

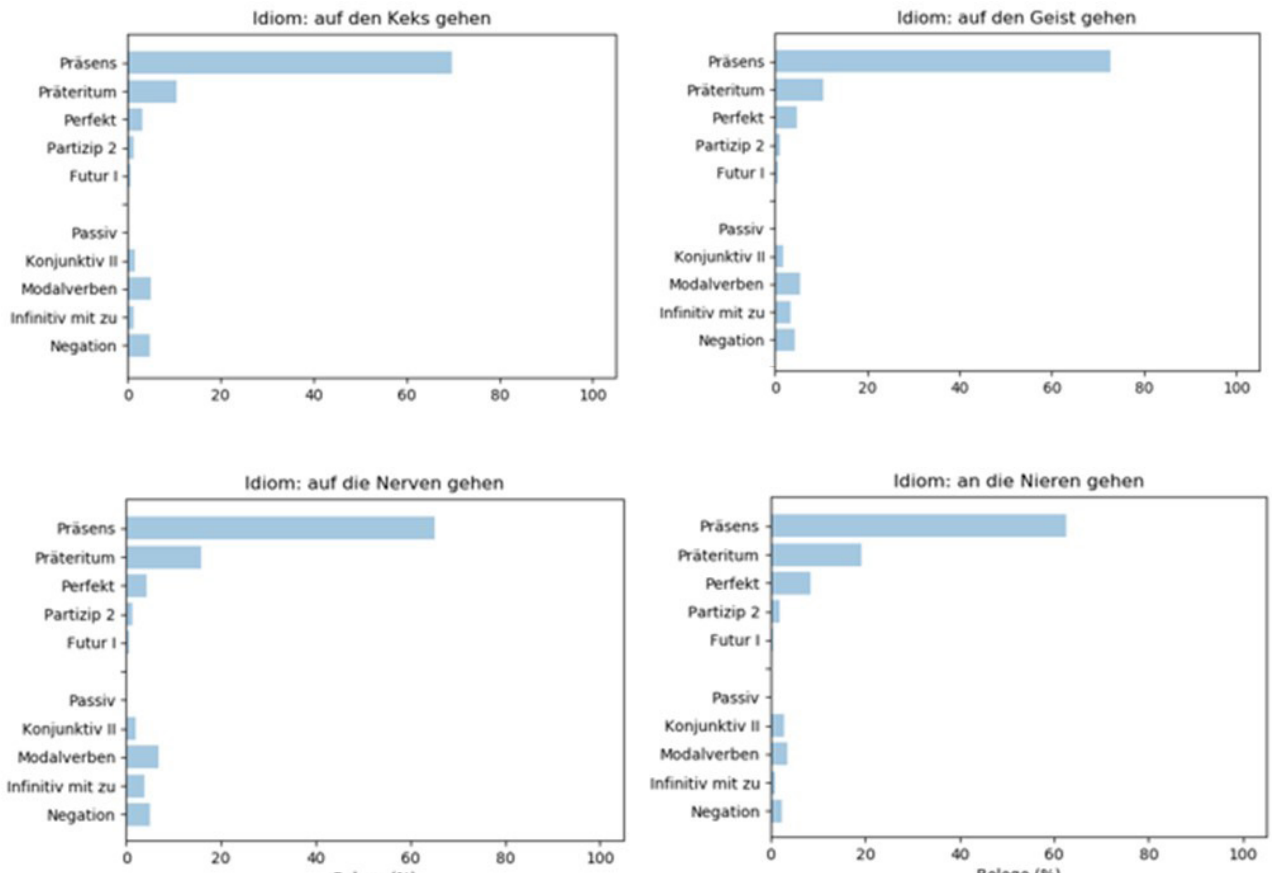

Figure 4: Comparison of graphs for *auf den Keks gehen, auf den Geist gehen, auf die Nerven gehen, an die Nieren gehen*

The group includes three idioms *auf den Keks gehen, auf den Geist gehen, auf die Nerven gehen* meaning ‘to get on sb.’s nerves’, as well *an die Nieren gehen* ‘to sadden someone’. They are mostly used in *Präsens* and less frequently in *Präteritum*. The usage in *Perfekt* compared to other tenses is rather rare. Though they have similar semantics and the identical verbal component *gehen*, there are still some differences in the frequency of usage with modal verbs, infinitive constructions and *Konjunktiv II*.

Other idioms belonging to the same group have quite different semantics: *auf der Stelle treten* ‘not to make any progress’, *auf Kohlen sitzen* ‘to be like a cat on hot bricks’, *ins eigene Fleisch schneiden* ‘to shoot oneself in the foot’, *dem Fass den Boden ausschlagen: Das schlägt dem Fass den Boden aus* ‘That’s outrageous’, *gegen den Strich gehen* ‘sth. rubs sb. the wrong way’, *Bude einrennen jmdm.* ‘to be overrun’, *ins Fäustchen lachen* ‘to laugh in one’s sleeve’, *aus der Reihe tanzen* ‘to break ranks’, *Armutszeugnis ausstellen* ‘to show sb.’s incapacity / incompetence / shortcomings’, *aus dem Sinn gehen, aus dem Kopf gehen: Es geht mir nicht aus dem Sinn [Kopf]* ‘It is always on my mind.’, *im Schilde führen* ‘to scheme sth.’, *vom Hocker reißen* ‘to knock sb.’s socks off’. A lot of them have negative connotations, but this is not unusual for the phraseological system generally, at least for the German language (Raykhshteyn 1980: 61). At the moment it is not clear why these particular idioms mostly occur in *Präsens*, and more idioms should be analyzed to draw more solid conclusions. The hypothesis is that the usage in a particular tense depends more on the pragmatics than on the semantics or the syntactical characteristics of the verbal components.

There is also the question whether graphs such as the ones used here provide the information about the difference between an idiom and free phrases. Let us compare the idiom *auf den Grund gehen* ‘to drill down on sth.’ with its two paraphrases *die Ursache finden*, *den Sachverhalt klären* and a free phrase with the same verb as in the idiom, *ins Bett gehen* ‘to go to bed’.

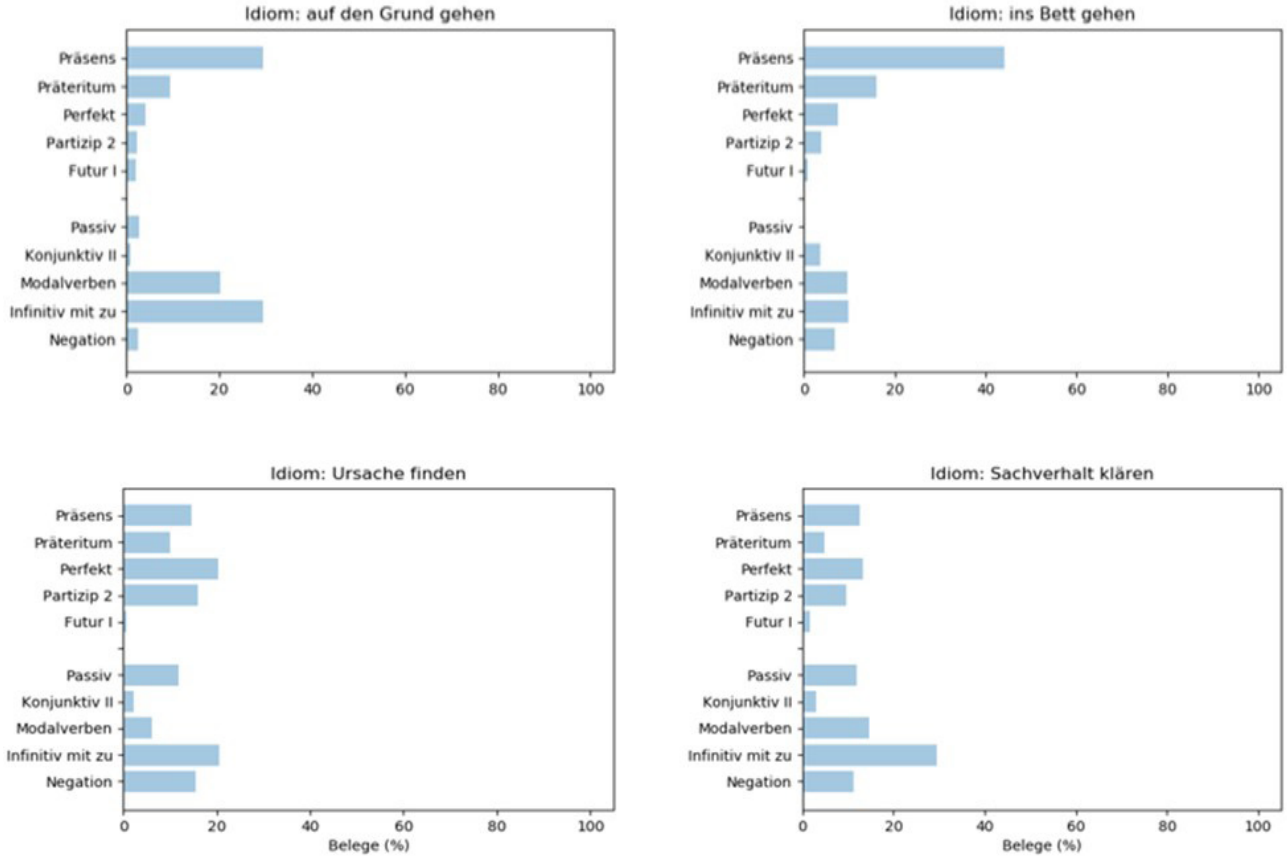

Figure 5: Comparison of graphs for *auf den Grund gehen, die Ursache finden, den Sachverhalt klären, ins Bett gehen.*

Unlike paraphrases and a free phrase containing the verb *gehen*, the idiom is used less often with negation. Compared with the paraphrases the idiom is also less often used in the passive, while the free phrase with the verb *gehen* almost never occurs in the passive, because the verb is normally used in active voice. Unlike paraphrases the idiom is more often used in *Präsens* and less often in *Perfekt*. Thus, the idiom's speech behavior corresponds neither with its paraphrases nor with the syntactic behavior of the verb used in it. On the one hand, the speech behavior of an idiom is determined by its semantics. On the other hand, the syntactic structure imposes its limitations.

6 Discussion

6.1 Comparison of Sketch Engine and DeReKo

The DeReKo is used in the project as a source of language material and not as a tool to search for all possible modifications. This is for the following reasons: only one DeReKo query should be made, and its processing takes up to several minutes, depending on the frequency of words in the query, the number of word forms and the complexity of the query. To obtain all possible information about modifications the researcher should make many queries and spend proportionally more time. The usage of the developed program for phraseological studies seems to be a more flexible option: it takes less time; the researcher does not have to formulate many queries; and frequency lists for all verb forms and for tokens preceding the nominal component can be easily obtained. Besides, it can be defined what tokens cannot appear between the verb and the prepositional phrase (e.g. commas for some modifications). This makes sense particularly for German, with its long sentences. Otherwise, the researcher gets a lot of contexts that do not contain the idiom, but only the components it consists of.

The reasons why Sketch Engine is not used are mostly the same. Though the biggest German corpora at Sketch Engine, German Web 2013, is half the size of DeReKo, it seems to be well-suited for studying idiomatic expressions, possibly because of its more colloquial character (e.g. for the idiom *nicht alle Tassen im Schrank haben* 'to not have all one's marbles' there are 1,595 occurrences in DeReKo[7] and 4,620 in German Web[8]). However, it does not allow users to export occurrences, and its query language is more difficult and less flexible than that of DeReKo, particularly if the target structures are phraseological units.

6.2 Application in Lexicographic Research

The results of the program can be applied in lexicographic research in the following ways. They can help to:

- write comments on acceptable modifications. For instance, if an idiom is a negative polarity item, we can explain in what type of contexts it can be used without negation;
- select illustrative material. The modifications profiles can help to find examples illustrating the usage of an idiom and do not contain rare modifications;
- specify the form of the dictionary entry, for example by answering the following questions:
  - Should a modal verb be a part of a dictionary entry? If so, which one?
  - Which article should be used in the dictionary entry? Example: *Den Ausschlag geben* 'to tip the balance'. In total, the program has found 8,215 idiom occurrences. Among these, 92% contain the definite article *den*. Other frequent lexical and syntactic modifications have also been found, e.g. *letzten* (1.1%) and *entscheidenden* (0.54%). Other tokens preceding the nominal component are less common, like *einen* (0.23%) and *keinen* (0.19%).

7 The query: Tassen /+w2 Schrank

8 The query: "Tassen" []{0,2} "Schrank" within <s/>

- Should the negation be a part of the dictionary entry?
  Example 1: *nicht aus dem Sinn gehen* 'not to go out of mind', 86% of occurrences contain the negation *nicht*.
  Example 2: *jmdm. kein Haar krümmen* 'not to lay a finger on sb.', 47% in the passive voice, 39% with the negation *kein*. Below the context from the DeReKo is provided, where the idiom is used without negation:

(1) Wir hatten noch Respekt vor den Lehrern, den meisten jedenfalls. Selbstverständlich gab es auch Lehrer, die wir nicht mochten – trotzdem hätten wir es nie gewagt, dem Lehrer auch nur ein Haar zu krümmen. (Braunschweiger Zeitung, 02.01.2006)?

7 Conclusion

Even if idioms possess comparable structures (verb plus prepositional phrase), they all have their own profiles of modifications. Such profiles cannot be inferred only from the semantics of an idiom, from the syntactical behavior of its verbal component or idioms' paraphrases. Ideally each idiom in a dictionary should be provided with a detailed description of its usage and modifications, as well as corresponding text examples. However, due to lack of space such dictionary articles would only be possible in electronic resources, but not in a printed dictionary.

To reduce the amount of work needed, a program developed as part of this project analyses and summarizes the acquired corpora data. There is a plan to expand the list of analyzed idioms to several thousands. This can be done after the first results have been thoroughly analyzed, and the program improved if needed.

References

Baranov, A.N., Dobrovol'skij, D.O. (2008). *Aspekty teorii frazeologii*. Moskva: Znak.

Deutsche Idiomatik. Accessed at: https://bitbucket.org/elena_krotova/deutsche_idiomatik [31/03/2018]

Deutsches Referenzkorpus. Accessed at: https://cosmas2.ids-mannheim.de/cosmas2-web/ [31/03/2018]

Dobrovol'skij, D. (2008). Idiom-Modifikationen aus kognitiver Perspektive. In Kamper, H., Eichinger, L.M. (Hrsg.) Sprache - Kognition -Kultur. Sprache zwischen mentaler Struktur und kultureller Prägung. Berlin / New York: de Gruyter, 2008, pp. 302-322.

Dobrovol'skij, D. (2006). Idiom dictionaries. In: Keith Brown, (Editor-in-Chief). Encyclopedia of Language and Linguistics. Second edition. Volume 5. Oxford: Elsevier, 2006, pp. 514-518.

Dobrovol'skij, D.O. (1997). *Nemetsko-russkiy slovar' zhivykh idiom*. Moskva: Metatekst.

Raykhshteyn, A.D. (1980) *Sopostavitel'nyy analiz nemetskoy i russkoy frazeologii*. Moskva: Vysshaya shkola.

Sketch Engine. Accessed at: https://www.sketchengine.co.uk/ [31/03/2018]