811.112.2

Кротова Елена Борисовна

Ученая степень: к.ф.н.

Ученое звание: -

Место работы: Институт Языкознания РАН, сектор германских языков

Должность: м.н.с.

Электронная почта: elena_krotova@inbox.ru

ИЗУЧЕНИЕ СИНТАКСИЧЕСКИХ МОДИФИКАЦИЙ ИДИОМ НЕМЕЦКОГО ЯЗЫКА НА ОСНОВЕ КОРПУСНЫХ ДАННЫХ[1]

Аннотация:

Статья посвящена корпусному подходу к изучению модификаций идиоматических выражений в немецком языке. Рассматривается только одна группа фразеологических единиц, а именно идиомы, которые характеризуются высокой степенью идиоматичности и стабильности. Для идиом характерна неполная парадигма: некоторые модификации, допустимые для свободных словосочетаний, недопустимы для идиом. В то же время, некоторые идиомы могут нарушать синтаксические ограничения, существующие для неидиоматических выражений. Несмотря на высокую степень идиоматичности, идиомы подвергаются разным модификациям (морфологическим, лексическим, лексико-синтаксическим и синтаксическим). Даже если идиомы обладают сопоставимой структурой (глагол плюс предложная группа), перечни допустимых для них модификаций могут значительно отличаться друг от друга. Это приводит к тому, что фразеографу необходимо составлять отдельно для каждой идиомы профиль допустимых модификаций, т.к. не

[1] Статья подготовлена при поддержке РФФИ, грант 18-012-00335.

представляется возможным делать обобщения. В идеале все модификации должны снабжаться соответствующими примерами употребления, полученными из корпусов. Такие словарные статьи, однако, были бы слишком объемными для печатного словаря и подходят только для публикации на электронном ресурсе.

Не все идиоматические выражения частотны, особенно в письменной речи, поэтому исследователю необходимо пользоваться крупными текстовыми корпусами, чтобы получить как можно больше примеров употребления рассматриваемой идиомы.

Для исследования модификаций идиом автор пользуется самым большим корпусом немецкого языка Deutsches Referenzkorpus (далее – DeReKo), содержащим более 42 млрд. токенов. Он является несбалансированным и состоит приблизительно на 95% из публицистики. Тем не менее, тот факт, что данный корпус является крупным и с его помощью можно получить тысячи примеров употребления идиомы в современных текстах, перевешивает его недостатки.

Автором была создана программа (на языке программирования *Python*), получающая информацию об употреблении идиом и их допустимых модификациях на материале корпусных данных DeReKo. Помимо этого, программа обобщает полученные данные в форме графиков. В статье будут подробнее рассмотрены возможности программы по получению информации об употреблении идиом в речи, а также каким образом полученные данные могут упростить работу фразеографа.

Ключевые слова: корпусная лингвистика, фразеография, модификации идиом

Перевод на английский

Elena Krotova, PhD

Institute of Linguistics, Russian Academy of Sciences, Department of Germanic Languages
Junior Research Fellow
Mail: elena_krotova@inbox.ru

CORPUS-DRIVEN ANALYSIS OF IDIOMS' SYNTACTIC MODIFICATIONS

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, the so-called idioms, which are characterized by a high degree of idiomaticity and stability. Idioms can possess an incomplete paradigm: It means that some modifications, which are possible in free phrases, are not always acceptable in idioms. On the contrary, some idioms can violate syntactic norms existing for non-idiomatic expressions. In spite of a high degree of stability, idioms still undergo different modifications (morphological, lexical, lexical-syntactic and syntactic ones).

Even if idioms possess comparable structures (verb plus prepositional phrase), they all have their own profile of modifications.

It means that for every idiom a phraseologist should write down its modification profile, because it is not possible to generalize. Ideally, all modifications should be provided with corresponding text examples. Due to lack of space such dictionary articles would only be possible in electronic resources, but not in a printed dictionary.

Not all idiomatic expressions are frequent, especially in the written speech. That is why if researchers want to get many text examples of one particular phraseme, they should make use of big corpora.

For the study of idioms' modifications the author has chosen the biggest corpus of German language is Deutsches Referenzkorpus (DeReKo) that contains more

than 42 billion tokens, but it is unbalanced and comprises up to 95% of newspaper texts. But still the fact, that the corpus is big and we can find hundreds of phraseme's occurrences in modern texts outweighs its unbalance. A Python-program has been created that obtains information about the usage of idioms and about their possible modifications from DeReKo. It also summarises the data in the form of graphs. The article will look further into the program opportunities to acquire information about idiom usage and in what ways such data can facilitate the work of a phraseologist.
Keywords: corpus linguistics, phraseography, modifications of idiomatic expressions

1. Введение

Создание словарных статей для фразеологических единиц, особенно идиом[2], является непростой задачей для фразеографа по ряду причин. Идиомы обладают сложной семантической структурой и синтаксическими особенностями, которые влияют на их употребление. Только имея подробное описание семантики и синтаксического поведения идиомы, не-носитель языка сможет верно употребить идиому в речи.

До появления крупных электронных корпусов составители фразеологических словарей опирались во многом на собственную языковую интуицию. Теперь же стало возможным верифицировать предоставляемую в словаре информацию и дополнить ее с помощью корпусных данных. В крупных корпусах можно найти тысячи примеров употребления отдельной идиомы. Этого количества примеров достаточно, чтобы подробно описать употребление идиомы в письменной речи.

[2] В данной работе под идиомами понимаются «сверхсловные образования, которым свойственна высокая степень идиоматичности и устойчивости» (Подробнее в [Баранов, Добровольский 2008: 57]).

Проблема, однако, состоит в том, что анализ такого большого количества данных отнимет у лингвиста слишком много времени, особенно если речь идет не о детальном анализе одной идиомы, а о составлении сотен статей для фразеологического словаря. Сократить время на мануальную обработку данных можно было бы с помощью компьютерной программы, которая проанализирует полученную из корпуса информацию, найдет различные модификации идиомы и сделает обобщения. В статье речь пойдет о создании такой программы для частотных идиом немецкого языка, а также будут представлены ее первые результаты.

2. Модификации идиом

Идиомы обладают высокой степень устойчивости, но, тем не менее, они могут подвергаться различным модификациям. На рис. 1 представлены графики с модификациями двух идиом немецкого языка:

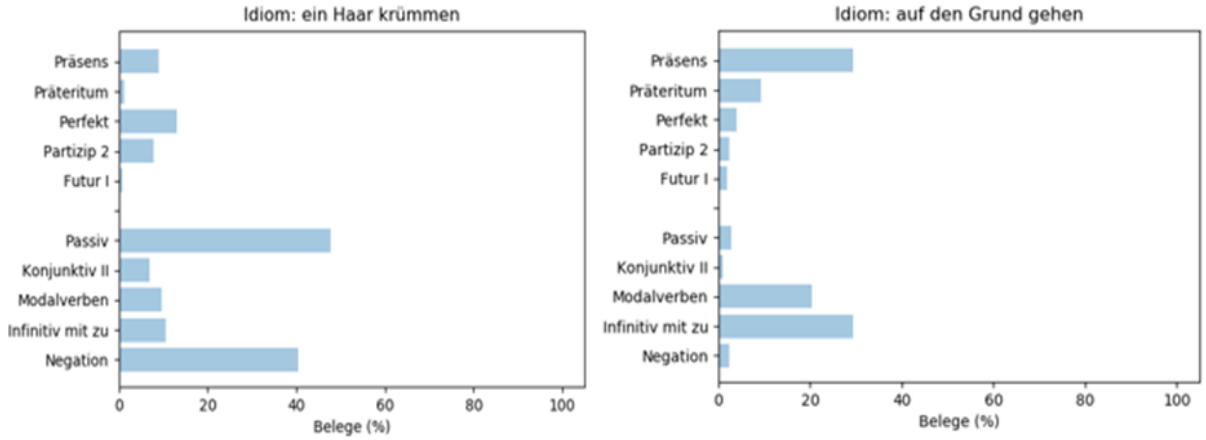

Рис. 1: идиомы *ein Haar krümmen jmdm.*, *auf den Grund gehen etw.*

Первая идиома *kein Haar krümmen jmdm.* ‘пальцем не тронуть кого-л.’ в основном используется в страдательном залоге и с отрицанием, в то время как для второй идиомы *auf den Grund gehen etw.* ‘докопаться до сути чего-л.’ такие модификации нехарактерны. Вторая идиома используется в основном в инфинитивных конструкциях с *zu*, в настоящем времени и с модальными глаголами.

В данной статье в основном рассматриваются синтаксические модификации, а также временные формы, в которых может употребляться глагольный компонент идиомы. Извлеченные из корпуса данные представлены на графиках. Другие типы модификаций (морфологические, лексические, лексико-синтаксические) анализируется только в некоторой мере.

В ходе исследования были проанализированы около ста идиом. Среди них минимальное количество допустимых модификаций было обнаружено только у идиомы *abwarten und Tee trinken* ‘подождем – увидим’, употребляющейся в основном в приведенной форме, и идиомы *dem Fass den Boden ausschlagen: Das schlägt dem Fass den Boden aus* ‘что-л. неслыханно; что-л. переходит все границы’. Эта идиома, в отличие от первой, может употребляться в разных временных формах, хотя такие случаи редки. Глагол также может менять свою форму в презенсе. Тем не менее, формы *schlägt aus* и *ausschlägt* составляют 86% от общего числа вхождений. Справа от идиомы представлен график для свободной фразы, содержащий глагольный компонент идиомы:

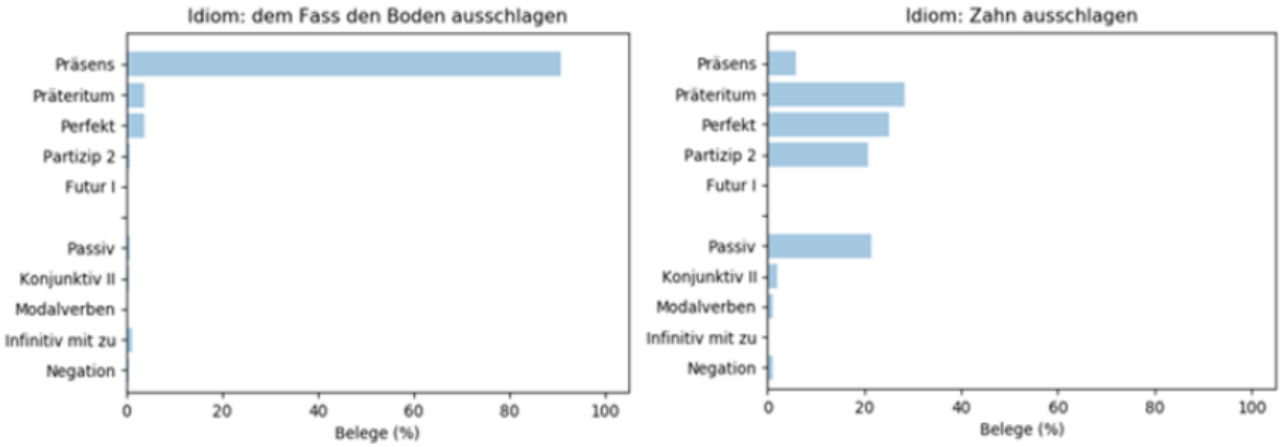

Рис. 3: сравнение идиомы *dem Fass den Boden ausschlagen* и свободного словосочетания *jmdm. einen Zahn ausschlagen* ‘выбить кому-л. зуб’

Как можно видеть, хотя идиома редко используется в страдательном залоге, ее глагольный компонент в свободном словосочетании встречается в страдательном залоге довольно часто. Структура идиомы сама по себе не

запрещает подобную модификацию, т.е. предложение *Dem Fass wurde der Boden ausgeschlagen* корректно с чисто грамматической точки зрения. Однако, такая модификация практически не встречается в речи. Более того, свободное словосочетание *den Zahn ausschlagen* редко встречается в презенсе, что, возможно, имеет прагматические основания: в прототипической ситуации данная фраза описывает результат физического действия, поэтому ее употребление в прошедшем времени более вероятно. Таким образом, лексикограф должен предоставлять пользователю словаря, изучающему язык, информацию о модификациях, которым подвергается идиома, т.к. о них нельзя судить на основании синтаксического поведения компонентов идиомы в свободных словосочетаниях.

3. Методология

Для анализа полученных данных используется программа, созданная автором статьи для проводимого исследования. Полученные данные представлены на сайте bitbucket.org [Deutsche Idiomatik]. На данный момент программа получает и анализирует следующую информацию:

- временные формы и словоформы, в которых употребляется глагольный компонент идиомы (*Präsens*, *Präteritum*, *Perfekt*, *Futur I*),
- синтаксические модификации, которым подвергается идиома, такие как использование в страдательном залоге, в *Konjunktiv II*; вместе с модальными глаголами и в инфинитивных конструкциях с *zu*. Глаголы (*werden* в страдательном залоге, глаголы в сослагательном наклонении, модальные глаголы) могут употребляться в любой временной форме. Такие контексты считаются только один раз и не учитываются в подсчете временных форм. К примеру, если идиома употреблена в страдательном залоге в настоящем времени, это вхождение считается только как случай употребления идиомы в страдательном залоге и не считается как случай употребления идиомы в настоящем времени.

Также программа ищет случаи употребления идиомы с отрицанием *nicht* или *kein*. Рассмотрим пример: для идиомы *jmdm. ein Armutszeugnis ausstellen: (jmd.) stellt (jmdm. mit etw. D) ein Armutszeugnis aus* ‘кто-л. считает кого-л. некомпетентным; кто-л. считает, что кто-л. расписался в своей (полной) несостоятельности’ было найдено 805 вхождений. Среди них 56% употреблений идиомы в презенсе, 14,66% в перфекте, 6,58% в форме причастия прошедшего времени (вспомогательные глаголы *sein* или *haben* не были найдены), 1,61% в страдательном залоге, 4,98% в сослагательном наклонении, 4,59% с модальными глаголами, 1,5% в инфинитивных конструкциях. Это составляет 99,72%. Кроме того, около 0,3% составляют случаи употребления в будущем времени.

Самой частотной словоформой для глагольного компонента рассматриваемой идиомы является *stellt aus* (28,9%). Идиома редко используется с отрицанием (4,59%). Самым частотным модальным глаголом, встречающимся с идиомой, является *können* (64% от всех вхождений с модальными глаголами).

Кроме того, анализируются токены, предшествующие именному компоненту идиомы. Например, существительному *Armutszeugnis* в 75% случаев предшествует артикль *ein*. Помимо этого, встречаются следующие варианты: определенный артикль *das* (1,49%), лексико-синтаксические модификации, такие как введение в структуру идиомы модификаторов *solches, dieses* (около 0,8% в каждом случае), прилагательных *politisches* (1,4%), *größeres, großes* (0,8% в каждом случае), *geistiges, eigenes* (0,6% в каждом случае). Таким образом, исследователь получает информацию о допустимых морфологических (*ein* или *das*) и лексико-синтаксических модификациях (*solches, dieses, politisches, größeres, großes*).

Программа также может искать предложения, содержащие вопрос, металингвистические конструкции и случаи, когда глагольный компонент

используется в первом лице, что может быть полезно при изучении случаев употребления идиомы в контекстах снятой утвердительности. Далее программа создает файлы, содержащие контексты употребления идиомы в разных временных формах и с разными модификациями. Для каждого случая создается отдельный текстовый файл, чтобы исследователь мог детально его проанализировать. Кроме того, программа обобщает полученные данные в форме графиков.

4. Применение во лексикографии

Результаты программы можно применять в лексикографических исследованиях следующим образом. С ее помощью можно:

- писать комментарии о допустимых модификациях. Например, если идиома является отрицательно поляризованной, можно указать, в каких типах контекстов она может быть употреблена без отрицания;
- выбрать иллюстративный материал. Профили модификаций могут помочь найти примеры употребления идиомы, которые хорошо иллюстрируют ее употребление в речи и не содержат редких модификаций;
- определить форму словарного входа, получив ответ, в том числе, на следующие вопросы:
  - должен ли модальный глагол быть частью леммы? Если да, то какой именно?
  - какой артикль должен быть употреблен в лемме?

Пример: *den Ausschlag geben* ‘иметь решающее значение, сыграть решающую роль‘. Всего программа нашла 8215 случаев употребления идиомы. Среди них 92% вхождений содержат определенный артикль *den*. Также были найдены следующие лексико-синтаксические модификации: ввод атрибутивных модификаторов *letzten* (1,1%), *entscheidenden* (0,54%).

Другие токены реже встречаются перед именным компонентом идиомы, например, *einen* (0,23%) и *keinen* (0,19%).

- Должно ли отрицание быть частью леммы?

Пример 1: *nicht aus dem Sinn gehen* ‘не идти из головы’. 86% от общего числа вхождений содержат отрицание *nicht*.

Пример 2: *jmdm. kein Haar krümmen* ‘пальцем не тронуть кого-л.’. В 47% случаев идиома употребляется в страдательном залоге, в 39% с отрицанием *kein*. Далее приводится контекст из DeReKo, в котором данная идиома употребляется без отрицания:

(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)

Даже если идиомы обладают схожей структурой (глагол плюс предложная группа), все они обладают своим собственным профилем модификаций. Подобные профили не могут быть выведены из семантики идиомы, из синтаксического поведения глагольного компонента идиомы или ее парафразов. В идеале каждая идиома в словаре должна снабжаться подробным описанием ее употребления в речи и допустимых модификаций с соответствующими иллюстративными примерами. Такие словарные статьи, однако, были бы слишком объемными для печатного словаря и подходят только для публикации на электронном ресурсе. Чтобы сократить объем требуемой работы, разработанная программа анализирует и обобщает полученные из корпуса текстовые данные. Планируется расширить список анализируемых идиом до нескольких тысяч. Это будет сделано после того, как первые результаты будут тщательно проанализированы и программа при необходимости доработана.

Список литературы:

1. Баранов А. Н., Добровольский Д. О. Аспекты теории фразеологии. М.: Знак, 2008. – 656 с.
2. Добровольский Д. О. Немецко-русский словарь живых идиом. М.: Метатекст, 1997.
3. Райхштейн А. Д. Сопоставительный анализ немецкой и русской фразеологии. М.: Высшая школа, 1980. – 142 с.
4. Deutsche Idiomatik. URL: https://bitbucket.org/elena_krotova/deutsche_idiomatik [31/03/2018]
5. Deutsches Referenzkorpus. URL: https://cosmas2.ids-mannheim.de/cosmas2-web/ [31/03/2018]
6. Dobrovol’skij D. 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.