наук. (10.02.04) / Порохницкая, Лидия Васильевна; государственный лингвистический университет. – Москва, 2004. – 195 с.

Qi Pan, A Tentative Study on the Functions and Applications of English Euphemism. – English Department, Zhenjiang Watercraft College, Zhenjiang, China: Theory and Practice in Language Studies, 2013.

Rawson, Hugh. How not to say what you mean. / Hugh Rawson. – Oxford University Press, 2002. – 501 pages.

Sketch Engine для лингвистических исследований

Е. Б. Кротова
Институт языкознания РАН (Россия)

Аннотация. Доклад является частью мастер-класса «Корпусные технологии в лексикографии (на материале германских языков)». Речь пойдет о корпусном менеджере и программе для анализа текста Sketch Engine, его функционале и применении в лингвистических исследованиях.

Ключевые слова: Корпусные технологии, корпусы, Sketch Engine

В данной статье представлена часть материалов мастер-класса «Корпусные технологии в лексикографии (на материале германских языков)». Речь пойдет о корпусном менеджере[6] и программе для анализа текста *Sketch Engine* [Sketch Engine], его функциях и применении в лингвистических исследованиях.

Под электронным корпусом в статье понимается «собрание текстов на данном языке, представленное в электронном виде и снабжённое научным аппаратом (разметкой)» [Плунгян 2005: 6]. Обе названные характеристики – электронная форма и наличие разметки – основное, что отличает электронные текстовые корпусы, с одной стороны, от традиционных текстовых корпусов, с другой стороны, от любого собрания текстов в электронной форме. Разметка представляет собой ту информацию, которая вносится в тексты при их обработке, и зависит от цели, с которой составляется корпус. К первичной разметке, присутствующей практически в каждом корпусе, относят токенизацию (разбиение на орфографические слова), лемматизацию (приведение словоформ к словарной форме) и парсинг (синтаксический анализ) (Подробнее в [Кротова 2013]).

Основной причиной, по которой для мастер-класса был выбран *Sketch Engine*, является наличие большого количества корпусов на данном ресурсе и широкий охват языков. Так, *Sketch Engine* предоставляет доступ к примерно 500 корпусам для более чем 90 языков. Полный список языков представлен по ссылке [Languages in Sketch Engine]. Для английского языка имеется 70 корпусов, среди них корпус, собранный из опубликованных в Интернете текстов, *English Web 2015* (enTenTen15), содержащий 15 млрд. слов (словоформ) и являющийся одним из самых крупных корпусов английского языка. Для немецкого языка имеется 17 корпусов, среди них веб-корпус

[6] Корпусным менеджером называется «специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме» [Захаров 2005: 3].

German Web 2013 (deTenTen13), содержащий 16,5 млрд. токенов[7]. Крупные корпусы немецкого языка собраны также в архивах Института немецкого языка в г. Мангейме (366 корпусов, разделенные на 18 архивов) [COSMAS II-Korpora]. Всего архивы содержат более 42 млрд. словоформ, из них 40 млрд. в открытом доступе. Тем не менее, поиск производится только по одному архиву, одновременно искать по всем архивам невозможно. Основной архив *W – Archiv der geschriebenen Sprache* содержит около 8 млрд. токенов и, таким образом, является меньше корпуса *German Web 2013*, представленного в *Sketch Engine*.

Размер корпуса принципиален в тех случаях, когда исследуются редкие явления языка, например, при изучении речевого поведения идиом. Чаще важен скорее состав корпуса и имеющаяся в нем разметка. Так, если исследуется синтаксис, важнее объема корпуса будет наличие богатой синтаксической разметки, ср. [Синтаксический корпус НКРЯ]. Корпусы, включающие разметку, которая делается полуавтоматически и требует проверки лингвистом, обычно на порядок меньше автоматически размечаемых корпусов (к примеру, размер Синтаксического корпуса – всего 1 млн. токенов). Кроме того, необходимо обращать внимание на состав корпуса. Для ряда исследований важно, чтобы корпус был сбалансированным[8]. С помощью сбалансированного корпуса можно получить представление о том, как изучаемый языковой феномен проявляет себя в целом в языке. Такими корпусами, в частности, являются Основной корпус НКРЯ (около 200 млн. словоформ), Британский национальный корпус[9] (около 100 млн., далее BNC) или *DWDS-Kernkorpus* (около 120 млн.). Большинство корпусов являются несбалансированными. Это могут быть корпусы, содержащий определенный тип текстов и/или создаваемые для изучения конкретного языкового явления. К примеру, *Sketch Engine* предоставляет доступ к корпусу медицинских текстов (*Medical Web Corpus*), к корпусу предлогов английского языка (*English Preposition Corpus*) и др. В зависимости от метаразметки[10] корпуса и возможностей корпусного менеджера можно также создать свой собственный подкорпус (виртуальный корпус). Так, при работе с BNC через *Sketch Engine* можно указать интересующий тип текстов, дату и место публикации, имя автора и т. п., и таким образом создать подкорпус, который будет лучше соответствовать цели проводимого исследования.

Для работы со *Sketch Engine* необходимо зарегистрироваться[11]. После регистрации пользователь выбирает язык и подходящий корпус. Далее будет

[7] В данной работе термин «токен» используется синонимично термину «словоформа», когда речь идет о размере корпуса.

[8] Сбалансированным считается корпус, который «содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т.п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода» [НКРЯ].

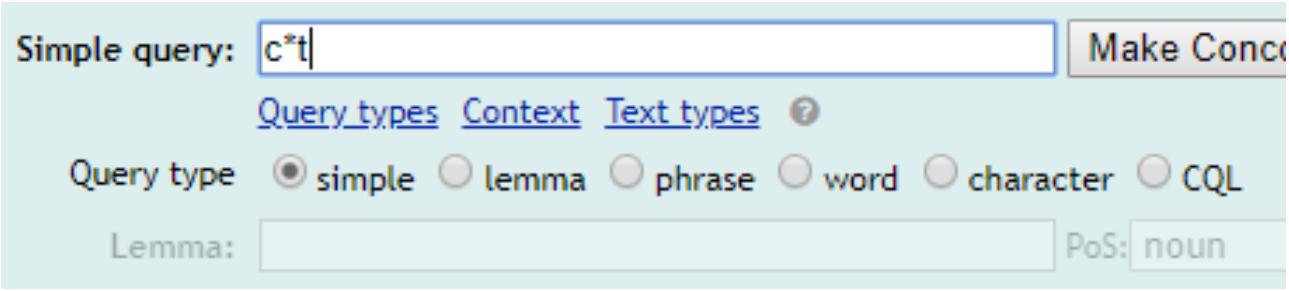
[9] Доступ к данному корпусу можно получить как через *Sketch Engine*, так и через сайт [Коллекция открытых корпусов], где собраны 14 корпусов английского языка.

[10] «Под метаразметкой понимается приписывание тексту атрибутов, характеризующих обстоятельства его создания, автора, тематику, жанровые особенности и др.» [НКРЯ].

[11] На первые 30 дней после регистрации предоставляется бесплатный доступ. Если пользователь по истечении месяца желает далее пользоваться ресурсом, ему необходимо оформить платную подписку.

кратко рассказано об основных функциях и некоторых возможностях применения *Sketch Engine* в лингвистических исследованиях. Подробную информацию о работе с данным корпусным менеджером можно найти в руководстве пользователя [User guide].

Поиск по корпусу может осуществляться с помощью простого поиска (по конкретной словоформе или по точной фразе), поиска по лемме, в том числе с указанием ее частеречной принадлежности, с помощью шаблонов (*wild cards*), а также с помощью специального языка (*CQL – Corpus Query Language*) для более сложных запросов. Поиск по шаблонам возможен в поле простого запроса (*Simple query*), но не в поле поиска по леммам (*Lemma*). При поиске с использованием шаблонов знак * заменяет 0 или более символов, а знак ? заменяет ровно один символ. К примеру, с помощью запроса “c*t” можно найти все словоформы, начинающиеся на *c*, заканчивающиеся на *t*. Так выглядит поисковый запрос:



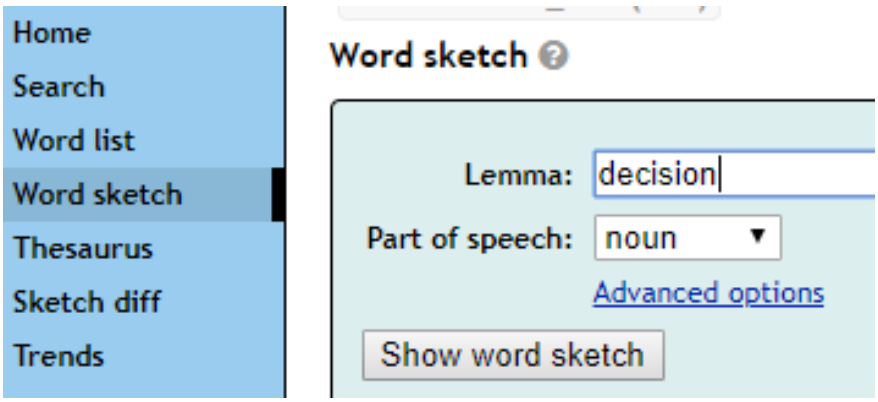
Полученные результаты можно отсортировать по разным параметрам, перечисленным слева от результатов поиска. Возможна сортировка по поисковому слову (*Sort – Node*), по левому и правому контексту, по источникам текста. Доступны данные по частотности конкретных словоформ (*Frequency – Node forms*) и встречающихся частей речи (*Frequency – Node tags*). Так, можно узнать, что наиболее частыми словоформами по выше приведенному запросу являются *cent*, *cost*, *court*, *cut* (поиск производился по корпусу BNC), а наиболее частыми частями речи существительные в ед. и мн. числе и прилагательные.

Более сложные запросы можно создавать с помощью [CQL]. Рассмотрим составление запроса для фразы *auf den Tisch schlagen «*ударить по столу»[12]. Запрос выглядит следующим образом: <s/> containing [lemma="schlagen"] containing ([lc="auf"][]{0,2}[lc="tisch"]). Он обозначает, что в рамках одного предложения (оператор <s/>) должны находиться лемма *schlagen*, а также словоформы *auf* и *tisch* на расстоянии не более 2 токенов друг от друга. При этом *schlagen* может находиться в любой позиции относительно *auf* и *tisch*. Подобный запрос к корпусам Института немецкого языка в г. Мангейме выглядел бы следующим образом: &schlagen /s0 (auf /+w2 Tisch), где оператор & обозначает лемму, оператор s0 – поиск в пределах одного предложения, оператор w2 – количество токенов между *auf* и *Tisch*. Как можно видеть, языки поисковых запросов могут сильно отличаться. Преимущество *Sketch Engine* в том числе в том, что один и тот же язык запросов можно применять

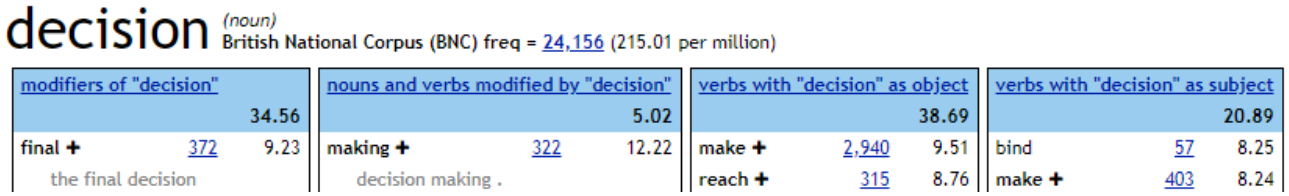
[12] Например, как часть фразеологизма *mit der Faust auf den Tisch schlagen* ‘стучать [бить] кулаком по столу (для наведения порядка)’.

к большому количеству разных корпусов и языков, без необходимости каждый раз при переходе к новому корпусу изучать и новый язык запросов, и знакомиться с новым интерфейсом.

Среди других корпусных менеджеров *Sketch Engine* во многом выделяет возможность создания так называемых скетчей (*word sketches*), которые и дали ему название. Под скетчем понимается описание речевого поведения слова, полученное автоматически путем обобщения информации о всех контекстах, в которых исследуемое слово встретилось в корпусе. Рассмотрим пример для существительного *decision*. Поиск по скетчам осуществляется в отдельном диалоговом окне:



В скетче подробно описывается речевое поведение слова из поискового запроса. Приводятся следующие частотные списки: модификаторы слова; существительные и глаголы, которые оно может модифицировать; глаголы, для которых оно может являться прямым дополнением и с которыми может выступать в функции подлежащего и т. д.:



При нажатии на число рядом со словом в списке (например, 372 для *final*) открываются контексты из корпуса, в которых *final* выступает модификатором *decision*. При нажатии на плюс рядом со словом *final* открывается так называемый мультискетч (*multiword sketch*), то есть скетч для обоих слов (*final decision*):

Для изучения синонимов удобно использовать сравнение скетчей (*word sketch differences*). Возьмем, к примеру, прилагательные *tasty* и *delicious* (*Sketch diff*). Программа сравнивает, в том числе, с какими модификаторами, существительными и глаголами могут использоваться оба слова. Результаты для модификаторов выглядят следующим образом:

modifiers of "tasty/delicious"	82	100	0.22	0.10
pretty	2	0	3.9	--
as	5	0	1.3	--
much	2	0	0.9	--
well	2	0	0.4	--
very	30	5	3.4	0.9
really	13	4	3.4	1.7
rather	3	3	2.6	2.6
so	6	18	1.3	2.8
quite	2	13	1.3	4.0
all	0	2	--	0.7
too	0	2	--	0.7
particularly	0	2	--	2.4
simply	0	4	--	3.2
especially	0	2	--	4.0
equally	0	3	--	4.3
truly	0	3	--	5.1
utterly	0	3	--	6.2
absolutely	0	13	--	6.5

Зеленым выделены слова, которые модифицируют *tasty*, но не могут модифицировать *delicious*. Белым – те, с которыми могут употребляться оба синонима. Красным – те, что характерны для *delicious*.

Кроме прочего, в *Sketch Engine* имеется автоматически создаваемый тезаурус (*Thesaurus*), с помощью которого можно искать синонимы для исследуемой языковой единицы, а также слова, употребляемые в похожих контекстах. Также для ряда корпусов подсчитано, частота каких слов в них сильно возросла или сократилась, какие новые слова возникли (*Trends*).

Выше представлены только основные возможности *Sketch Engine*, более подробную информацию можно найти в [User guide]. С помощью данного корпусного менеджера можно работать с большим количеством корпусов для разных языков и проводить разнообразные лингвистические исследования, как синхронные, так и диахронные, на различных языковых уровнях.

Литература

Захаров В. П. Корпусная лингвистика / Учебно-методическое пособие. СПб., 2005. – 48 с.

Коллекция открытых корпусов. URL: https://corpus.byu.edu/

Кротова Е. Б. Корпусная фразеография (на материале немецкого языка): диссертация... канд. филол. наук: 10.02.04. М, 2013. – 296 с.

НКРЯ. Национальный корпус русского языка. – URL: http://ruscorpora.ru/corpora-intro.html

Плунгян В. А. Зачем нужен Национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003 – 2005. М.: Индрик, 2005. С. 6-20.

Синтаксический корпус НКРЯ. URL: http://ruscorpora.ru/search-syntax.html

COSMAS II-Korpora. URL: http://www.ids-mannheim.de/cosmas2/projekt/referenz/archive.html
CQL. Corpus Query Language. URL: https://www.sketchengine.eu/documentation/corpus-querying/
Deutsches Referenzkorpus. URL: https://cosmas2.ids-mannheim.de/cosmas2-web/
DWDS. DWDS-Kernkorpus. URL: https://dwds.de/r
Languages in Sketch Engine. URL: https://www.sketchengine.eu/user-guide/user-manual/corpora/by-language/
Sketch Engine. URL: https://www.sketchengine.eu/
User guide (Sketch Engine). URL: https://www.sketchengine.eu/user-guide/

Оценочная категоризация в паремиологии (на материале русских и немецких народных примет)

М. А. Кулькова
Казанский (Приволжский) федеральный университет (Россия)

Аннотация. Настоящая статья посвящена изучению языковых средств выражения положительной оценки в паремиях русского и немецкого языков с точки зрения когнитивно-дискурсивного подхода.

Ключевые слова: положительная оценка, оценочность, народная примета, русский язык, немецкий язык

Целью настоящей статьи является выявление общих и различных характеристик положительных оценочных явлений в народных приметах русского и немецкого языков с точки зрения современной когнитивно-дискурсивной парадигмы языкового знания.

Устно-поэтическая природа народных примет предопределяет понимание данного уникального типа паремий в качестве устойчивых высказываний неопределенно-референтного типа, в которых запечатлен богатейший когнитивный опыт народа, конденсирующий результаты предыдущих этапов когнитивной деятельности – эмпирического познания окружающего мира и понятийного осмысления полученной информации носителями той или иной лингвокультуры [Кулькова 2011: 349], [Kulkova 2015: 356].

Согласно Н. Д. Арутюновой, общеоценочное прилагательное *хороший* (так же, как и прилагательное *плохой*), «имеют обобщающее, конденсирующее значение» [Арутюнова 1984: 17]. Как отмечает А. Н. Баранов, «общие оценки представляют собой метаоценочную категорию, сферой действия которой является не просто ситуация или соответствующая ей пропозиция, а своеобразная «амальгама», сплав из описания положения дел и приписанных этому положению дел частных оценок различных типов» [Баранов 1989: 78]. Семантика оценочного слова *хорошо / gut* помимо оценочной модальности имплицирует в себе множество других модальных смыслов: долженствование (должен), необходимость (надо), совет (хорошо бы), рекомендация (следует) и т.д., формируемых в «прагматическом периметре» высказывания. Сюда следует в первую очередь отнести контекстуальное окружение оценочного высказывания, участников коммуникативного акта, пресуппозиции продуцента и реципиента высказывания. Например: *Скоро хорошо не родится, Как паутина полетит – хорошо сеять, Рыба будет ловиться особенно хорошо, если все*