Адамович Светлана Васильевна
УО «Гродненский государственный университет имени Янки Купалы»
Республика Беларусь, Гродно
sv_adam@mail.ru

Использование электронной базы данных немецкого языка Deutscher Wortschatz в лингвистических исследованиях и лингводидактике

Проект Deutscher Wortschatz представляет собой универсальный банк данных немецкого языка, комплексный электронный мегаресурс, сочетающий возможности корпуса текстов, лексикографических баз данных и on-line грамматик. Его использование при изучении разнообразных языковых явлений открывает перед исследователем неограниченные возможности, существенно повышая объективность полученных результатов. Объем лексикона составляет более 9 миллионов словарных статей и около 160 миллионов предложений (более 2 миллиардов словоупотреблений). Следует отметить также возможность свободного доступа к базе данных лексикона посредством сети Интернет для использования в научных и образовательных целях.

В структурном плане словарная статья лексикона существенно отличается от словарных статей типичного толкового словаря. Здесь представлена разнотипная лингвистическая информация об искомом слове: количество его употреблений в лексиконе и класс частотности, морфологический состав слова, его грамматические данные: принадлежность к определенной части речи (для склоняемых и спрягаемых слов указываются формы флексий), синонимический ряд и семантическая группа, в состав которых входит искомое слово. В словарной статье приводятся также словосочетания и контексты, равные или превышающие предложение, которые иллюстрируют функционирование слова в речи, его дистрибуцию. Их количество колеблется в зависимости от класса частотности слова.

Представленные в словарных статьях морфологические, синтаксические, семантические, прагматические, статистические данные можно использовать при проведении исследований в об-

ласти семантики, для изучения синонимии, антонимии, гипонимии, для проведения различных грамматических и морфологических анализов, а также в процессе обучения немецкому языку как иностранному.

Базу данных Deutscher Wortschatz мы использовали в качестве источника материала исследования при изучении средств выражения семантической категории аппроксимации. В результате был сформирован репрезентативный корпус немецкоязычного материала исследования, который составили 13056 контекстов с аппроксиматорами. Полученный исчерпывающий языковой материал послужил основой для выделения 156 разнообразных средств выражения аппроксимации, выявления особенностей их семантики в языке и специфики функционирования в речи.

Анализ лингвистической информации словарных статей аппроксиматоров и предложений с аппроксиматорами из корпуса текстов лексикона способствовал всестороннему изучению лексических средств выражения семантической категории аппроксимации в немецком языке, а также послужил основой для сопоставительного исследования аппроксиматоров в немецком, русском и белорусском языках.

Таким образом, использование компьютерных языковых ресурсов позволяет автоматизировать этап сбора фактического материала исследования, обработать репрезентативные выборки и проанализировать разнотипный языковой материал большого объема.

Бачаева Саглар Егоровна
Калмыцкий институт гуманитарных исследований РАН
Республика Калмыкия, Элиста
basaeg@mail.ru

О принципах составления «Толкового словаря языка калмыцкого героического эпоса «Джангар»

В современной лексикографии накоплен огромный опыт по составлению различного рода словарей: этимологических, исторических, синонимических, орфоэпических, обратных, диалектологических, толковых. Наиболее известные, полные и более доступные толковые словари: «Словарь современного русского литературного языка», «Большой толковый словарь русского языка», «Толковый словарь русского языка» С.И. Ожегова и Н.Ю. Шведовой, «Современный толковый словарь русского языка» Т.Ф. Ефремовой и др. В Калмыцком институте гуманитарных исследований Российской академии наук ведется работа по созданию Толкового словаря языка калмыцкого героического эпоса «Джангар».

Толковый словарь – это единое цельное издание, обладающее всеми признаками текста, который объясняет лексическое значение слова и определяет все его значения. Порядок расположения слов в словаре алфавитный и алфавитно-гнездовой. При составлении толковых словарей одной из важнейших задач является правильное описание и толкование слов, дача им ясной, полной, четкой формулировки, понятной для читателей. В нашем толковом словаре используется описательный и описательно-синонимический способы семантических определений, если слово нельзя будет описать этими способами, то будут применяться другие способы, т.е. в словаре допускаются разновариантные комбинации толкований.

Во введении – описательной части словаря – дается его краткая характеристика, назначение, структура словарной статьи, содержание, принципы пользования. Главная базовая часть словаря – словарная статья. В структуру словарной статьи в разрабатываемом словаре входят следующие элементы: заголовочное

слово, абсолютная частота употребления, транскрипция, которая поможет пользователю словаря правильно произнести слово, грамматические и стилистические пометы, объяснение их значений – толкования и примеры использования в данном значении слов, фразеологизмы, дериваты, сложные слова, коллокации. В словник включаются все слова, встречающиеся во всех циклах эпоса «Джангар».

Все заголовочные слова приводятся в начальной форме: для имен существительных это форма именительного падежа, единственного числа, для глаголов – форма причастия в будущем времени.

Иллюстративные примеры сопровождается ссылками, откуда они взяты, в квадратных скобках указывается источник данного словоупотребления, в котором приводится название версии эпоса «Джангар» в сокращенной форме, через двоеточие номер песни – римскими цифрами, например: [ЭО: V], [МБ: I], [БЦ: II]. Примеры отделяются один от другого точкой с запятой.

Принципы составления толкового словаря языка калмыцкого героического эпоса «Джангар» разработаны на основе анализа теоретической базы составления существующих толковых словарей. Данный словарь может быть полезен людям, как владеющим калмыцким языком, так и начинающим его изучать.

Беленчикова Ренате
Университет им. Отто фон Герике
Германия, Магдебург
renate.belentschikow@ovgu.de

Большой Русско-немецкий словарь (РНС) как опыт объять необъятное

1. Концепция любого словаря определяет его предмет и функции, круг его пользователей, его словарную базу и объём. Проект РНС предполагает охватить максимум объёма русского языка. Словарь издаётся с 2003 г. при Майнцской Академии наук и литературы, а его концепция была разработана ещё в конце 1980-х годов. Как преимущественно рецептивный словарь, РНС должен помочь немецкоязычным пользователям-профессионалам (филологам, переводчикам) в раскрытии русскоязычных текстов всех жанров, от классической русской литературы XIX века до наших дней. Ситуации пользования словарём предполагаются как в сферах двуязычной коммуникации, так и в научных поисках. Более того, РНС благодаря объёму может служить основой для других, менее объёмных словарей, например, учебных.

2. Названным выше концептуальным задачам соответствуют количество заголовочных слов (250 тыс.) и широкая словарная база. Помимо лексического ядра литературного языка в словник РНС отбирается специальная лексика, которая является релевантной в общей коммуникации (медицина, спорт, компьютер, интернет и т.д.), а также историзмы и устаревшие слова, лексемы из социолектов и просторечия и даже обсценизмы, т.е. пласты лексики, составляющие по русской лексикографической традиции отдельные словари. Кроме того, в словарных статьях РНС приводятся фразеологизмы и многословные неидиоматические наименования.

3. В зависимости от данных параметров разработана абстрактная микроструктура РНС, в которой главная роль отводится подбору эквивалентов и объяснению семантики и прагматики лексем языка-источника. В связи с этим большую значимость приобретает прозрачная для пользователя филиация исходного

слова. К исходным словам (их лексико-семантическим вариантам) даются грамматическая характеристика, пометы употребления, особенности произношения, литературные грамматические, акцентологические и орфоэпические варианты, обязательная сочетаемость заголовочных слов, а также, в качестве иллюстративного материала, свободные словосочетания. Несмотря на рецептивный характер РНС, для круга адресатов словаря данная информация оказывается востребованной и облегчает восприятие текстов и нормативую оценку употребляемых языковых средств. В частности, крайне важной является прагматическая характеристика лексем и вариантов при помощи дифференцированной системы функциональных и стилистических помет. При этом, все заголовочные слова должны характеризоваться в соответствии с абстрактной микроструктурой каждой части речи по единым параметрам.

4. Обычно рецептивный печатный словарь отличается стремлением к расширению словника при ограничении объёма информации о заголовочных словах. Для РНС это ограничение было снято, так как он с самого начала был запланирован как электронный словарь, хотя небольшим тиражом (200 экз.) издаётся и печатный вариант. С 2010 года словарь разрабатывается в режиме онлайн в редакционной системе ABBYY® Lingvo™ Content. На основе банка данных создаются как печатная, так и электронная версия словаря для публикации в интернете. В настоящее время словарные материалы первых томов вводятся в электронный банк данных, и тем самым обеспечивается их сохранение для возможных запросов в будущем.

Беляева Лариса Николаевна
Российский государственный педагогический университет
Россия, Санкт-Петербург
lauranbel@gmail.com

Переводная лексикография как основной ресурс технологического процесса перевода

Современный уровень развития лингвистических технологий определяет необходимость уточнения места и функций технического перевода и самого технического переводчика в особой технологической цепочке, включающей системы машинного перевода, комплекс автоматизированных словарей, предметно ориентированный корпус текстов, комплекс прикладных программ. Оставив в стороне рассуждения о том, является перевод ремеслом или искусством, мы просто вынуждены определить, каковы функции переводчика и терминолога в новой структуре, как должен быть организован обмен информацией в технологической цепочке перевода.

Особую часть в технологической цепочке и средствах лингвистической поддержки перевода составляют лексикографические ресурсы, ориентированные на необходимость выполнения терминологической работы: для терминолога существует насущная необходимость реагировать быстро (и стандартным образом) для того, чтобы удовлетворять требования к обработке информации и выделять не зарегистрированные ранее или просто новые терминологические единицы. Различия самих исходных текстов, уровней специализации текстов, целей и профилей конечных пользователей и уровня автоматизации объясняют отсутствие универсальных методов для решения задачи извлечения терминов из текстов. Результаты работы терминолога должны вводиться в систему лексикографических ресурсов до того, как переводчик получает текст и результат машинного перевода. В современной технологической цепочке перевода терминологическая работа не просто является самостоятельным звеном, но осуществляется до собственно перевода.

Современные многоязычные лексикографические ресурсы по степени универсальности и доступности можно разделить на государственные (например, поддерживаемые Комиссией ЕС) и инициативные, разрабатываемые корпорациями или исследовательскими группами.

Банк данных EuroTermBank представляет собой один из самых мощных государственных терминологических банков, охватывая все языки Европейского союза и латынь. В этом лингвистическом ресурсе объединено 133 локальных ресурса, разработанных в различных бюро перевода ЕС, в нем 2 650 976 терминов (число постоянно увеличивается), 710 705 словарных статей, 221 512 дефиниций на 33 языках. Структура информации в базе данных EuroTermBank предполагает различные опции выбора исходного языка и языка перевода, выбора предметной области, выбора формы представления информации. При выборе конкретных опций предоставляется информация о вариантах перевода в различных предметных областях и о зафиксированных в базе данных словосочетаниях.

Ресурс базы терминов EuroTermBank может рассматриваться как опробованная модель многоязычного сетевого ресурса, создание которого является, безусловно, актуальным как для языков национальных республик России, так и для языков таможенного союза, поскольку может обеспечить корректную терминологическую и лексикографическую поддержку для перевода документов в различных областях сотрудничества и знаний.

Задача оперативного пополнения переводных лексикографических ресурсов решается на основе автоматизации процесса извлечения терминов из параллельных или сопоставимых текстов. Методы извлечения терминов варьируются от независимого от конкретного языка извлечения n-граммов с использованием оценки относительной частоты и степени терминологичности до лингвистически обоснованных методов на основе синтаксического анализа и применения моделей терминологических словосочетаний. Комбинация статистических и лингвистически обоснованных приемов является наиболее удачным подходом в практических инструментальных средствах создания и ведения лексикографических ресурсов.

Бешенкова Елена Виленовна
Институт русского языка им. В.В. Виноградова РАН
Россия, Москва
evbeshenkova@gmail.com

Проблемы определения нормы и ее соотношения с узуальными предпочтениями в словаре «Слитно или раздельно: написание слов с отрицанием *не*»[1]

Орфографические словари подразделяются на два типа: прескриптивные (предписывающие определенную орфографическую норму-кодификацию) и дескриптивные (описывающие либо узус во всем охвате, либо только устоявшуюся часть узуса, т.н. «узуальную норму»). Разрабатываемый словарь представляет синтез этих двух подходов. В нем, с одной стороны, отражается «узуальная, объективная» норма современного письма, а с другой стороны, в случае устойчивой узуальной вариативности отмечаются два варианта, но при этом один из них отмечается как предпочтительный.

При выборе рекомендаций необходимо учитывать следующие внутрисистемные факторы, а также факторы, внешние по отношению к системе письма.

1. Предназначение письма – обеспечить письменную коммуникацию современников и понимание текстов предшествующих поколений.

2. Коммуникативная значимость, целесообразность орфографических единиц оценивается с позиции читающего и пишущего под диктовку.

3. Письменная форма должна быть средством различия «грамотный» – «неграмотный», а не «свой» – «чужой», «модный» – «немодный».

4. Наличие внутрисистемных законов:

1) системное противопоставление общеотрицательной частицы *не* и приставки *не*;

[1] Работа выполнена при поддержке гранта РГНФ № 14-04-00444 «Теоретические и лексикографические проблемы написания слов с отрицанием *не*».

2) наличие контекстов, в которых это противопоставление нейтрализуется, т.е. в которых написание определяется не системой, а нормой;

3) формирование аттракторов и нестабильной зоны в саморазвивающейся системе;

4) сохранение традиции, историческая устойчивость письма.

5. Наличие и степень устойчивости стихийной, узуальной нормы.

При выборе стратегии нормирования кодификатор учитывает все приведенные факторы, которые не выстраиваются в иерархию, могут обусловливать одинаковые или противоположные результаты, и руководствуется поставленной целью (обеспечение письменной коммуникации и сохранение системы письма) и способом протекания процесса управления системой письма (в эпохи социальной стабильности предпочитается плавный, адаптационный характер протекания процесса плавный, а в эпоху реформ возможен и скачкообразный способ).

Таким образом, узуальные предпочтения являются одним из пяти факторов, учитываемых при принятии кодификации. Если узуальная норма однозначна, то она принимается и как кодифицированная, даже если это написание противоречит системе (напр. *не беден* имеет приставочное значение, но пишется только раздельно *он весьма не беден*). Если написание вариативно, то это отмечается в словарях, но в словаре может быть одна из двух помет: *предпочтительно слитно* или *чаще раздельно* (при сильном преимуществе раздельного варианта). Решение рекомендовать слитное написание при равноправности вариантов в узусе принято в результате выявления исторической тенденции к увеличению доли слитного написания.

Бешенкова Елена Виленовна
Институт русского языка им. В.В. Виноградова РАН
Россия, Москва
evbeshenkova@gmail.com
Иванова Ольга Евгеньевна
Институт русского языка им. В.В. Виноградова РАН
Россия, Москва
olliva95@yandex.ru

Реализация принципов объяснительных словарей в «Объяснительном русском орфографическом словаре-справочнике»

«Объяснительный русский орфографический словарь-справочник» принадлежит объяснительному типу орфографических словарей. В нем соединены два принципа построения объяснительного словаря: исторический и синхронный.

Книга состоит из двух частей: собственно словаря и справочника – сборника правил, переработанных в соответствии с задачами словаря. Как способ соотнесения словаря и справочника в словарной статье используется и номер параграфа в справочнике, и название орфограммы, совпадающее с названием параграфа в правилах (*непроизносимая согласная, глаголы на ова(ева)/ыва(ива), приставка пре/при*) либо являющееся частным случаем данной орфограммы (*корень сед/сид*). Кроме того, часто, а в случаях, когда написание слова нельзя обосновать современным правилом, обязательно, дается историческая справка. Если современное написание слова не соответствует ни правилам, ни происхождению, его написание трактуется как закрепившееся. В словарной статье предусмотрена зона комментария, в которую помещается различная дополнительная информация.

Справочник по орфографии содержит правила, дополненные и переработанные в соответствии с особенностями лексикографической формы представления информации (нормы написания при этом не затрагиваются) и описания большого словарного массива. С учетом обоих этих факторов была выстроена логика соподчинения правил, были сформулированы некоторые новые

правила (прежде всего, в разделе слитных-дефисных-раздельных написаний), были дополнены и закрыты списки исключений из правил.

«Объяснительный русский орфографический словарь-справочник» не просто дает ответ на прямой вопрос, почему существует то или иное написание. Благодаря тесной связи словарной и справочной частей книги возникает приращенное знание, показывающее системную организацию русской орфографии, современную картину существующих в ней связей и закономерностей.

Пример словарной статьи:

массм**е́**д**иа** (англ. mass media от лат. māssa «массовый, масса» + media мн. ч. от medium «средство, посредничество») см. масс; *одиночная*/*двойная согласная в сложносокращенных словах и графических сокращениях*: **исключение** – сохранение группы согласных в сложносокращенном слове § 11 п. 2 искл.; *слитно*/*дефисно*/*раздельно*: пишется **слитно** как сложносокращенное существительное (*массовые медиа*) § 44 или *слитно*/*дефисно*/*раздельно*: закрепившееся **слитное** написание сложного существительного с не употребляющейся самостоятельно первой частью на согласную § 46 п. 2 ◊ Несмотря на то, что слово является прямым заимствованием, в настоящее время оно все больше воспринимается как сокращение сочетания *массовые медиа*, напр.: *в массовые медиа периодически проникают дикие, по европейским меркам, скандалы*; *традиционные массовые медиа* (*и прежде всего – телевидение*) *останутся основными локомотивом медиаиндустрии и рекламного рынка*; *массовые медиа будоражит информация о…* См. коммент. к *масс*.

Блинова Ольга Владимировна
Санкт-Петербургский государственный университет
Россия, Санкт-Петербург
0973000@gmail.com

Слово в транскрипте и словаре (конвенции токенизации для корпуса «Один речевой день»)[1]

Токенизация (выделение минимальных линейных компонентов) является типовым этапом обработки письменного текста. При токенизации текста, оформленного согласно стандартной орфографии и пунктуации, решается проблема отделения слов, скобок, кавычек, знаков пунктуации, манипуляции алфавитно-цифровыми комплексами, сокращениями и т. д. Такая работа, которая к тому же может быть выполнена машинными средствами, отличается от задач выработки приёмов ручной сегментации, сообразно которым может действовать коллектив расшифровщиков в ходе создания корпуса устных текстов.

Звуковой корпус повседневного общения «Один речевой день» формируется с применением методики непрерывной записи речи информантов-добровольцев. Основой корпуса являются расшифровки (транскрипты) звукозаписей, подвергающиеся многоуровневой разметке (в том числе автоматической).

С начала работы над корпусом выработаны конвенции оформления транскриптов. Рекомендации для расшифровщиков требуют, чтобы цепочки символов, которые считаются отдельными токенами, были с обеих сторон отделены пробелами. В частности, таким образом оформляются: словоформы; символы, обозначающие паралингвистические явления, паузы, наложение речи и др.; знак членения на синтагмы (/), знак конца фразы (//), другие знаки фразовых границ (?, !). Для многословных выражений введено одно правило: названия из двух или более слов объединяются с помощью подчёркивания (например, *Грачи_прилетели*).

Относительно написаний, в графическом представлении которых задействован дефис, говорится следующее:

[1] Исследование выполнено в рамках проекта РНФ № 14-18-02070 «Русский язык повседневного общения: особенности функционирования в разных социальных группах».

*-*то*, *-*таки* даются как отдельные слова, за исключением неопределённых местоимений и местоимённых наречий (*где-то*, *куда-то* и т. д.), а также слова *всё-таки*.

Сложные слова с дефисом (за исключением прилагательных, обозначающих цвет, имён собственных и некоторых наречий, имеющих в своей основе дуплеты), а также лексические повторы принимаются за два отдельных слова.

Для достижения единообразия графического представления однотипных вхождений необходимо сформировать дополнения к принятым конвенциям. Такие дополнения, по-видимому, должны содержать:

1. Правила, касающиеся категорий слов, у которых сохраняется дефисное орфографическое написание. Таким образом оформляются: сложные предлоги; неопределённые местоимения, местоимённые наречия с *кое-, -то, -либо, -нибудь*; наречия с *по-* (*по-моему*), существительные с *пол-* (*пол-ложки*); прилагательные, образованные по модели типа *светло-жёлтый, англо-русский*; имена собственные; вхождения, компоненты которых не встречаются по отдельности (*давным-давно*) и др.

2. Правила, касающиеся вхождений, для которых дефис в транскрипте заменяется на пробел, подразумевают: раздельное написание различных «дублетов»: прилагательных (*быстрый быстрый, синий синий*), наречий (*долго долго*), глаголов (*сидел сидел*), междометий (*ай ай*, *ха ха*); раздельное написание частиц, в том числе ‘*ка*’, ‘*таки*’, ‘*де*’, существительных типа ‘*сестра близнец*’, числительных со значением приблизительного количества типа ‘*один два*’ и др.

3. Списки многословных выражений, объединяемых подчёркиванием: составные предлоги, составные союзы, названия, формы взаимного местоимения *друг_друга*, формы типа *кое_у_кого* и др.

Информация о написаниях, обобщённая с учётом употребительности той или иной категории вхождений, может быть сформулирована с использованием мнемонических приёмов; кроме того, для упрощения работы расшифровщика внутри среды ELAN могут использоваться так называемые контрольные словари.