Т. А. Архангельский
(факультет филологии НИУ ВШЭ)

Корпуса албанского, калмыцкого, лезгинского и осетинского языков

В статье представлены четыре электронных корпуса, созданные в 2011 году в рамках Программы фундаментальных исследований РАН «Корпусная лингвистика»: албанский, калмыцкий, лезгинский и осетинский. Даётся описание интерфейса и функциональности этих корпусов, освещаются технические вопросы, которые пришлось решать при их создании, обсуждаются перспективы их развития. Особое внимание уделяется вопросам составления грамматических словарей и автоматической грамматической разметки корпусов.

1. Введение

Лингвисту при изучении какого-либо языкового объекта или явления — лексемы, грамматической категории, конструкции и т. д. — часто бывает необходимо получить большое количество примеров этого явления в исследуемом языке. Традиционный способ получения языковых примеров — опрос информантов — имеет ряд очевидных недостатков, например, невозможность моментально получить ответ от информанта в любое время, необходимость опроса большого количества информантов для получения информации о языке в целом, а не об идиолекте одного человека; и, возможно, самое главное — неспонтанность получаемых данных, отрефлексированный вариант языка, данные не столько о реальном использовании языка, сколько о языке в представлении информанта. Для решения этих проблем лигвистами в последние два десятилетия активно создаются специальные электронные инструменты — корпуса языков.

Корпусом языка называется собрание текстов на этом языке, в котором текстам или их фрагментам (абзацам, предложениям, словоформам или даже морфемам) приписана дополнительная лингвистически релевантная информация (аннотация) и которое снабжено поисковым механизмом, позволяющим производить поиск по этой информации. Аннотация может включать в себя любую информацию, в зависимости от задач, стоящих перед создателями корпуса и исследователями. При этом корпус, в отличие от электронной библиотеки, предназначен в первую очередь не для просмотра полных текстов имеющихся в нём произведений (хотя в некоторых корпусах такая возможность имеется). Основная задача корпуса — предоставить исследователю языка возможность быстро получать реальные языковые примеры по заданному запросу и выяснять относительную частоту появления в текстах языковых объектов, соответствующих этому запросу. То, какие поисковые запросы позволяет делать корпус, зависит от того, какая дополнительная информация в него внесена.

Среди информации, которой могут снабжаться тексты корпуса, следует в первую очередь выделить метатекстовую информацию и пословную разметку. Метатекстовая информация — это информация, которая приписывается тексту в целом. Она может включать в себя имя автора, время создания текста, размер текста и т. п. (например, ср. описание метатекстовой разметки Национального корпуса русского языка в [Савчук 2005]). Пословная разметка, т. е. дополнительная информация, указываемая при каждой словоформе текста, содержит прежде всего грамматическую информацию, но также может включать перевод на другой язык, информацию о тех или иных семантических признаках словоформы и др. Грамматическая информация может включать в себя лемму (начальную форму), часть речи, словоклассифицирующие и словоизменительные грамматические признаки.

В то время как для многих крупных языков Европы корпуса существуют и давно используются при проведении лингвистических исследований (например, английские

British National Corpus <http://www.natcorp.ox.ac.uk/> и Corpus of Contemporary American English <http://corpus.byu.edu/coca/>, Национальный корпус русского языка <http://www.ruscorpora.ru/>, Чешский национальный корпус <http://ucnk.ff.cuni.cz/>, Венгерский национальный корпус <http://corpus.nytud.hu/mnsz/index_eng.html> и другие), для большинства менее распространённых языков, в том числе для большинства языков России, корпусов не существует. В 2011 г. в рамках программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» были созданы, среди прочего, корпуса албанского, калмыцкого, лезгинского и осетинского (иронский диалект) языков. Эти корпуса имеют схожую функциональность, поскольку созданы на одной платформе и с использованием одних и тех же принципов и одной и той же компьютерной системы грамматической разметки. В настоящее время корпуса размещены на сайтах ossetic-studies.org (осетинский) и web-corpora.net (остальные); все корпуса доступны по ссылкам со страницы <http://web-corpora.net/>. В разработке албанского корпуса принимали участие М. С. Морозова, М. В. Домосилецкая, А. Ю. Русаков и Е. Д. Бернацкая, калмыцкого — А. Э. Ванькаева, лезгинского — Д. С. Ганенков, осетинского — О. И. Беляев и А. П. Выдрин.

В разделе 2 настоящей статьи будет дана общая характеристика созданных корпусов с описанием их функциональности. В разделе 3 будет кратко описан процесс создания корпуса и особо рассмотрена его важная часть — создание системы морфологического анализа. В последнем разделе будут изложены перспективы дальнейшей работы над этими корпусами.

2. Характеристика корпусов

Все упомянутые корпуса включают письменные тексты на нормативном варианте соответствующего языка. Приблизительные объёмы корпусов таковы: албанский — 750 тыс. словоупотреблений, калмыцкий и лезгинский — по 800 тыс., осетинский — 5 млн. На данный момент эти корпуса уступают в объёме корпусам крупных европейских языков (например, объём Национального корпуса русского языка в настоящий момент составляет почти 200 млн. словоупотреблений), и добавление новых текстов является одним из приоритетов их развития. Однако стоит учесть, что корпуса письменных текстов на малых и средних языках не могут достичь таких показателей, как корпус, например, русского языка, по той причине, что такого количества текстов на этих языках просто не существует. В корпуса включены созданные в XX—XXI вв. художественные произведения, мемуары, очерки и (в осетинском и калмыцком) некоторые периодические издания.

Все корпуса снабжены базовой метатекстовой и подробной грамматической разметкой. Метатекстовая информация включает в себя название текста, имя автора или название периодического издания, год написания (или временной промежуток, если точная дата неизвестна или написание текста заняло более одного года) и жанр. Грамматическая разметка включает в себя начальную форму слова и информацию о всех грамматических категориях, выраженных в данной словоформе морфологически. Грамматическая разметка производилась программными средствами (см. ниже) и имеется более чем у 70% словоформ в каждом корпусе. При подобной автоматической разметке одна и та же словоформа может получить несколько разных грамматических разборов (например, албанское *nga* может быть предлогом, союзом или одной из двух форм глагола *ngas*). Снятие омонимии, т. е. выбор для каждой такой словоформы в каждом её контексте в корпусе одного из предложенных разборов, является довольно сложной и времяёмкой задачей и в данных корпусах не проводилось, поэтому любая подобная словоформа во всех контекстах будут иметь все присовенные ей разборы (ср. подкорпус со снятой омонимией в Национальном корпусе русского языка).

Все корпуса используют одну и ту же платформу (поисковую систему и веб-интерфейс), которая изначально была создана компанией Corpus Technologies для

Восточноармянского национального корпуса <http://www.eanc.net/> [Даниэль et al. 2009]. Эта платформа обладает мощным поисковым функционалом, позволяющим задавать сложные контекстные запросы, использовать метаинформацию для формирования пользовательских подкорпусов (то есть подмножества текстов, в которых осуществляется поиск, — например, только периодические издания или только тексты до 1980 г.) и настраивать вид получаемых результатов по многим параметрам.

Основным инструментом поиска является запрос по словоформе, лемме и грамматической информации. При выборе вкладки «Словоформа» будут найдены все вхождения введённой словоформы, т. е. все контексты, где встречается введённая в поисковое поле последовательность символов. При выборе вкладки «Лемма» будут найдены все вхождения всех словоформ, которые были размечены как формы лексемы, лемма которой совпадает с введённой последовательностью символов. Таким образом, при поиске по лемме могут быть найдены только те словоформы, которым был приписан грамматический разбор, в отличие от поиска по словоформе. При задании искомой словоформы или леммы можно использовать знак «*», означающий любое количество любых символов, знак дизъюнкции «|», знак конъюнкции «&» и знак отрицания «~». Например, по запросу «mi*&~mirë» в албанском корпусе будут найдены все вхождения всех словоформ, начинающихся на mi-, но не совпадающих со словоформой «mirë» («добро»), а по запросу «*тты|*тыл» в осетинском корпусе будут найдены все словоформы, оканчивающиеся на -тты или -тыл. В случае использования нескольких логических операций их порядок можно задавать при помощи скобок. Все специальные знаки, включая звёздочку, можно использовать в запросе неограниченное число раз, однако высокая сложность запроса может привести к существенному увеличению времени поиска, которое в случае запроса обычной сложности составляет несколько секунд.

Кроме поиска по лемме и словоформе, возможен поиск по переводу словоформы на другой язык, что удобно для исследователей, не владеющих исследуемым языком на достаточном уровне. Размеченные словоформы в албанском корпусе снабжены переводом на английский язык, в калмыцком и лезгинском — на русский, а в осетинском — как на английский, так и на русский. При поиске по переводу можно использовать те же знаки, что и при поиске по лемме/словоформе.

В поле «Грамматика и части речи» можно задать запрос на грамматические признаки словоформы. Функциональность поиска по грамматическим категориям в целом аналогична функциональности Национального корпуса русского языка. В данное поле помещаются краткие обозначения грамматических категорий, которые можно как ввести с клавиатуры, так и выбрать из списка. В поле грамматических категорий, как и в поле ввода леммы/словоформы, можно использовать знаки трёх логических операций, комбинируя наборы грамматических показателей искомых словоформ. Запрос на грамматические признаки можно комбинировать с запросом на лемму или словоформу. Так, если в калмыцком корпусе ввести «*д» в поле «Лемма» и «N,ins» в поле «Грамматика и части речи», будут найдены все вхождения существительных в творительном падеже, начальная форма которых заканчивается на -д.

Существует также ряд дополнительных условий, которые можно наложить на искомую словоформу или её ближайший контекст, раскрыв группу полей под заголовком «Дополнительно». С помощью этих полей можно потребовать наличия какого-либо знака пунктуации слева и/или справа от словоформы или отсутствия какой бы то ни было пунктуации, ограничить позицию словоформы в предложении (в начале, в конце, в середине), регистр словоформы (с прописной буквы, все прописные, все строчные, как введено), а также разрешить или запретить находить словоформы, имеющие несколько омонимичных разборов.

Поиск можно одновременно проводить по нескольким словоформам, определённым образом расположенным друг по отношению к другу в контексте — иными словами,

искать словосочетания и конструкции. Для этого, задав нужное количество запросов на словоформы (по умолчанию в панели создания запроса включены поля для двух словоформ, но их количество можно увеличивать), необходимо указать расстояния между вхождениями этих словоформ в тексте. Расстояния можно указывать в виде диапазона, а также использовать отрицательные значения. Например, если в лезгинском корпусе указать в грамматике первой словоформы «N», в грамматике второй — «A» и установить расстояние между ними от -1 до 1, будут найдены все сочетания «прилагательное + существительное» и «существительное + прилагательное». Вместо указания точного числового диапазона в словах можно потребовать, чтобы искомые словоформы находились в одном предложении, в т. ч. на любом расстоянии, или в одном документе (с указанием диапазона расстояний в предложениях).

Выдача результатов возможна в нескольких представлениях: полном, кратком, KWIC (Key Word In Context) и глоссированном. Основной особенностью представления KWIC является то, что ключевая словоформа каждого найденного предложения отцентрована. В глоссированном формате под каждой словоформой текста подписаны её начальная форма и грамматические характеристики (при этом разбиение словоформы на морфемы не производится). Имеется возможность настроить количество результатов на одной странице выдачи и порядок сортировки примеров (по умолчанию примеры выдаются в случайном порядке). В лезгинском, осетинском и калмыцком корпусах можно выбрать алфавит выдачи: кириллица или транслитерация.

При выводе результатов поиска показываются сами найденные примеры, их общее количество и количество документов, в которых они были найдены. При каждом примере указывается название текста, автор или название периодического издания и год создания текста. У каждой словоформы, имеющей хотя бы один грамматический разбор, во всплывающем окошке указывается лемма, грамматические признаки, разделённые на словоклассифицирующие и словоизменительные, и перевод. Каждый пример содержит одно предложение текста; контекст можно расширять. Максимальный размер расширенного контекста составляет 7 предложений (3 предложения слева от найденного, само найденное предложение и 3 предложения справа). Контексты большего размера не предоставляются по юридическим причинам: многие тексты корпуса являются объектом авторского права, и размещение больших фрагментов текста могло бы быть признано его нарушением.

Документы, по которым производится поиск, можно ограничить с помощью опции «Задать подкорпус». В частности, можно ограничить подкорпус текстами определённых жанров, авторов, лет или просто выбрать тексты из списка.

Интерфейсы всех корпусов выполнены на русском и английском языках, а интерфейс албанского корпуса переведён также на албанский язык.

3. Морфологический анализ

Создание корпуса включает несколько основных этапов. На первом этапе необходимо собрать тексты, включаемые в корпус. Источниками текстов могут быть Интернет, издательства, отсканированные книги и расшифровки устной речи. В рассматриваемых корпусах использовались первые два источника. На втором этапе необходимо внести в собранные тексты грамматическую информацию. В случае корпусов небольшого размера (десятки тысяч слов) это можно сделать вручную. Однако для более крупных корпусов этот вариант практически неприемлем, поскольку требует очень больших ресурсов. Поэтому для разметки текстов необходимо разработать систему автоматического морфологического анализа. Третьим этапом является создание или адаптация существующей поисковой платформы и вывеска корпуса в Интернете.

Разработка системы морфологического анализа является одним из центральных этапов создания корпуса. Задача создания такой системы для данного языка состоит из нескольких частей. Первым шагом является описание морфологической системы языка,

т. е. описание имеющихся в нём частей речи, грамматических значений и способов их выражения, описание классов лексем с одинаковым словоизменением (напр., склонений существительных) и т. п. Для рассматриваемых языков эта задача была выполнена экспертами — разработчиками соответствующих корпусов. Вторым шагом является создание грамматического словаря, включающего в себя словоизменительные таблицы и перечень лексем с указанием их словоизменительного типа, основы/основ и всей остальной информации, необходимой для построения полной парадигмы. Примером такого словаря является созданный ещё в 70-е годы и многократно переиздававшийся Грамматический словарь русского языка [Зализняк 2007]. Для рассматриваемых языков, однако, подобных словарей не существовало. Третьим шагом является создание парсера — программы, способной размечать текст, используя данные грамматического словаря.

В ходе работы над проектами их участниками были созданы электронные грамматические словари рассматриваемых языков (на основе существующих грамматик, переводных словарей и других материалов). На данный момент словари носят пилотный характер и содержат от 4 до 15 тысяч лексем.

В качестве парсера во всех корпусах использовалась разработанная автором настоящей статьи система UniParser. Эта система одинаково подходит как для флективных языков, какими являются албанский и частично осетинский, так и для агглютинативных, какими являются лезгинский и в особенности калмыцкий. Система UniParser принимает на вход грамматический словарь в специальном формате, основанном на YAML, и использует его для разметки текстов. Кроме того, она может по запросу порождать полные парадигмы имеющихся в словаре лексем (листинг), что удобно для проверки правильности заполнения словаря и присваивания словоизменительных классов.

Грамматический словарь, предназначенный для использования в этом парсере, состоит из нескольких файлов, самыми важными из которых являются список словоизменительных типов, список лексем и список продуктивных дериваций. Список словоизменительных типов предназначен для описания изменяющейся части слов, в то время как список лексем содержит основу или основы каждой лексемы со ссылкой на её парадигму и другую лексическую информацию.

Каждый словоизменительный тип имеет название и представляет из себя перечисление флексий (морфем или комбинаций морфем), покрывающее всю парадигму. При каждой флексии указывается набор выражаемых ею грамматических значений. Если лексемы соответствующего словоизменительного типа имеют несколько основ, то при флексии указывается номер основы, с которой она употребляется (нумерация начинается с нуля). Ниже приведён фрагмент описания одной из именных парадигм албанского языка, в котором перечислено несколько флексий множественного числа (неопределённый номинатив: пустая флексия, неопределённый генитив: *-ve*, определённый номинатив: *-t*, определённый генитив: *-ve*), каждая из которых употребляется с первой основой лексемы:

```
-paradigm: 5a
 -flex: <0>.
  gramm: indef,pl,nom
 -flex: <0>.ve
  gramm: indef,pl,gen
 -flex: <0>.t
  gramm: def,pl,nom
 -flex: <0>.ve
  gramm: def,pl,gen
```

Хотя в рассматриваемых языках подавляющее большинство грамматических значений выражается аффиксами справа от основы или изменениями в основе, формат описания словоизменения подходит и для языков, в которых грамматические значения выражаются

аффиксами слева от основы, полиаффиксами (составными аффиксами, имеющими несколько несмежных частей, в том числе располагающихся с обеих сторон от основы), трансфиксами и т. п. В случае флективных языков с небольшими парадигмами (албанский) в каждом типе перечисляются все флексии. Для облегчения работы с агглютинативными языками, такими как калмыцкий, где глагольная парадигма состоит из сотен форм, а на границах аффиксов практически не происходит контактных изменений, существует возможность задать парадигму комбинацией нескольких элементарных подпарадигм. Так, парадигму калмыцкого глагола можно представить в виде комбинации последовательно идущих подпарадигм, одна из которых включает в себя только показатели каузатива, другая — только аспектуальные показатели, третья — только лично-числовые показатели и т. п. — своего рода аналог «грамматики позиций».

Каждая запись, содержащая информацию о лексеме, содержит лемму, основу или несколько основ, название словоизменительного типа, перевод и любую другую лексическую информацию. Лексема может иметь несколько ссылок на словоизменительные типы (например, в албанском грамматическом словаре парадигмы множественного и единственного числа были описаны как разные парадигмы, и у каждого существительного, соответственно, отдельно указывался его словоизменительный тип в единственном и во множественном числе). Основ также может быть несколько; основы разделяются знаком «|». Как указывалось выше, в случае, если лексема имеет несколько основ, дополнительно распределённых в парадигме, при каждом аффиксе соответствующего словоизменительного типа должен быть указан номер основы, с которой он употребляется. Ниже приведено описание лексемы *vajzë* «девочка» из абланского словаря:

```
-lexeme
 lex: vajzë
 stem: vajza.|vajzë.|vajz.
 gramm: N,f,anim
 paradigm: 3a
 paradigm: 5a
 transl_en: girl
```

Эта лексема имеет три основы. В описании также указаны грамматическая информация (существительное, женский род, одушевлённое), словоизменительный тип (парадигма 3а содержит флексии единственного числа, парадигма 5а, часть которой приведена выше, — множественного), начальная форма и английский перевод. Каждая основа лексемы представляет собой слово, в котором точками обозначаются места присоединения аффиксов, в то время как в аффиксах точка используется для обозначения частей основы (то есть своего рода «точек сцепления» основы и показателя). При построении словоформы парсер соединяет нужные основу и аффикс, вставляя на место точек в основе нужные части аффикса и наоборот. Например, определённый номинатив множественного числа данной лексемы будет образован сложением первой основы лексемы с соответствующим аффиксом: vajza. + .t → vajzat. Формат предусматривает также специальные средства для описания вариативности основ и аффиксов.

Одной из важных особенностей системы UniParser является возможность описания продуктивных деривационных схем. Деривационная схема представляет собой правило, по которому из имеющейся в словаре лексемы можно автоматически получить другую лексему. В деривационной схеме описываются изменения, вносимые в основу, лемму и грамматическую информацию исходной лексемы и даётся ссылка на словоизменительный тип деривированной лексемы. При этом лексические признаки деривированной лексемы могут как замещать лексические признаки исходной лексемы, так и дополнять их, а лемма

деривированной лексемы может быть образована по заданным правилам от её основы или совпадать с леммой исходной лексемы.

Перечисленные особенности позволяют использовать деривационные схемы для двух типов задач. В первом случае с их помощью автоматически порождаются и заносятся в грамматический словарь новые лексемы. Например, в осетинском языке имеется продуктивный способ образовать прилагательное от существительного с помощью суффикса *-он*. После занесения соответствующей информации в список деривационных схем с указанием, что подобная схема может применяться ко всем существительным, словарь пополняется прилагательными на *-он*, образованными от каждого имеющегося в словаре существительного, что позволяет не заносить такие прилагательные в словарь в виде отдельных записей и тем самым существенно экономить усилия по составлению грамматического словника. Другой, несколько более сложный пример такого рода из осетинского словаря — образование перфективной формы глагола с помощью префикса *ных-* — приведён ниже:

```
-deriv-type: V-nyx
 lex: <0>ных[.]ын
 stem: ных[.].
 regex-stem: x[^ъ].*
 gramm: +pv,pv-ny
```

Точкой в квадратных скобках обозначается основа деривируемого глагола, точка без скобок имеет то же значение, что и в описании лексемы, — место присоединения словоизменительных аффиксов. В этой деривационной схеме указано, что начальная форма деривированной лексемы образуется присоединением префикса *ных-* и инфинитивного суффикса *-ын* к первой основе исходной (деривируемой) лексемы; основа или несколько основ деривированной лексемы образуются присоединением префикса *ных-* к основам исходной лексемы. К грамматическим параметрам исходной лексемы добавляются значения pv («глагол с превербом») и pv-ny («глагол, содержащий один из алломорфов преверба *ны-*»), что необходимо для обеспечения возможности поиска в корпусе глаголов с превербами. В поле regex-stem с помощью языка регулярных выражений задано условие, ограничивающее область действия только теми лексемами, основа которых начинается на *х-*, но не на *хъ-*. В ходе обработки грамматического словаря парсер, применяя это деривационное правило, например, к глаголу *хауын* «падать; иметь отношение», автоматически добавляет в словарь его перфективный коррелят *ныххауын*.

Во втором случае деривации можно использовать для описания наборов форм, которые традиционно интерпретируются как принадлежащие к одной парадигме. Пример такого типа в русском языке — причастные формы глаголов. Хотя традиционной грамматикой причастие считается формой глагола, с точки зрения экономичности описания удобнее представлять его в виде регулярного отглагольного прилагательного, задав соответствующую деривационную схему. Такое представление позволяет избежать описания всех причастных форм в каждой глагольной парадигме. При этом деривированные формы можно считать принадлежащими парадигме исходной лексемы — для этого достаточно указать в деривационной схеме, что лемма деривированной лексемы совпадает с леммой исходной (таким образом, при поиске по этой лемме будут найдены все формы: и исходные, и деривированные). Такой тип дериваций был использован, например, в калмыцком корпусе для описания субстантивного словоизменения причастий.

Получив на вход словарь и грамматику языка, описанные в указанных файлах, парсер строит базу данных, на основе которой в дальнейшем можно производить морфологическую разметку текстов. При разметке парсер принимает на вход текстовый файл; результатом работы является файл на XML-языке, приближенном к используемому

в Национальном корпусе русского языка. Ниже приведён фрагмент размеченного албанского текста:

```
   <w><ana lex="hedh" gr="V,3,sg,aor,ind,act"
transl_en="throw"></ana>hodhi</w>

   <w><ana lex="sy" gr="S,m,inanim,def,pl,acc"
transl_en="eye"></ana><ana lex="sy" gr="S,m,inanim,def,pl,nom"
transl_en="eye"></ana>sytë</w>
```

Внутри тэга <w> заключена одна словоформа и грамматическая информация о ней. Последняя содержится в атрибутах тэга <ana>, например, lex — начальная форма, gr — перечисленные через запятую грамматические пометы, transl_en — перевод на английский язык. Если парсер присвоил лексеме несколько альтернативных разборов, словоформа будет содержать несколько тэгов <ana>, как во второй словоформе из примера.

4. Перспективы развития

Одной из главных целей дальнейшей работы над корпусами является увеличение текстовой базы и расширение грамматических словарей.

При сборе новых текстов важен не только их объём, но и тип. Чтобы корпус был репрезентативным, т. е. действительно мог представлять язык во всех его вариантах, в нём должны быть представлены тексты разных стилей и жанров: художественные и нехудожественные; переводные и непереводные; поэзия и проза; пресса, официальные документы, тексты из интернет-форумов и т. д., причём желательно, чтобы ни один из типов текстов не имел бы существенного количественного преобладания над остальными. Ясно, что не для всех перечисленных языков тексты всех этих жанров существуют, но стремиться разнообразить состав корпуса несомненно следует.

Расширение грамматических словарей позволит повысить качество грамматической разметки. В настоящее время грамматические словари для перечисленных языков позволяют разбирать не менее 70% встречающихся словоупотреблений. Если для достижения этого результата хватало словарей, имеющих не более 15000 лексем, то для дальнейшего повышения качества грамматической разметки это количество необходимо будет существенно увеличить. Поскольку в целом лексемы заносятся в словарь в порядке убывания их частотности, каждая следующая внесённая лексема будет иметь всё меньшее влияние на общее количество разобранных словоформ в корпусе; таким образом, каждый следующий процент прироста количества разобранных словоформ требует всё большего количества дополнительных лексем в словаре. Для пополнения словаря можно пользоваться автоматически порождаемым частотным списком неразобранных словоформ, внося в словарь соответствующие лексемы в порядке частоты их появления в корпусе.

Другим направлением дальнейшей работы является расширение функциональности корпуса. В частности, планируется в число вариантов отображения результатов поиска добавить глоссированную выдачу (в настоящее время есть вариант «псевдоглоссированной» выдачи — текст с грамматическим подстрочником, но без разбиения на морфемы). Формат описания грамматики уже сейчас позволяет парсеру выделять в словоформах аффиксы, остаётся добавить в интерфейс возможность отображать эту информацию. Ещё одним возможным расширением функциональности является размещение на страницах корпусов статистической информации, такой как частотные списки словоформ и лексем (ср. соответствующую страницу на сайте Восточноармянского национального корпуса).

Наконец, кроме расширения имеющихся четырёх корпусов, в рамках продолжающейся корпусной программы Президиума РАН планируется создание на той же платформе

новых корпусов. На ближайший год запланировано создание крупного корпуса новогреческого языка и ряда корпусов средних языков России и мира.

5. Библиография

Даниэль et al. 2009 — Даниэль М. А., Левонян Д. В., Плунгян В. А., Поляков А. Е., Рубаков С. В., Хуршудян В. Г. Восточноармянский национальный корпус. // Армянский гуманитарный вестник. Вып. 2/3-II. М.—Ереван: Зангак-97, 2009, стр. 9—33

Зализняк 2007 — Зализняк А. А. Грамматический словарь русского языка: Словоизменение. 4-е изд., испр. и доп. — М.: Русские словари, 2007

Савчук 2005 — Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003—2005. Результаты и перспективы. — М., 2005, стр. 62—88