Восточноармянский национальный корпус www.eanc.net

Eastern Armenian national corpus www.eanc.net

Хуршудян В. Г. (vk@corpustechnologies.com),
Даниэль М. А. (misha.daniel@gmail.com),
Левонян Д. В. (dl@renovacapital.com),
Плунгян В. А. (plungian@gmail.com),
Поляков А. Е. (pollex@mail.ru), **Рубаков С. В.** (rubakov@gmail.com)

Corpus Technologies, Москва

Восточноармянский национальный корпус (ВАНК) — это лингвистическая информационно-поисковая система, основанная на обширной коллекции текстов (около 110 млн.) на восточноармянском языке, покрывающая период с середины 19-го века до наших дней и снабженная мощной и гибкой поисковой функциональностью. ВАНК находится в открытом доступе в интернете (www.eanc.net).

ВАНК является репрезентативным, сбалансированным и полным электронным корпусом современного восточноармянского языка. Проект ВАНК был запущен в январе 2006 г. по инициативе группы московских исследователей и компании CorpusTechnologies. Летом 2007 г. был открыт портал www.eanc.net, на котором был размещен первый релиз корпуса. Второй релиз был размещен на том же портале весной 2008 г., а третий релиз — в марте 2009 г.

Третий релиз отличается от предыдущих объемом поискового корпуса (около 110 млн. словоупотреблений вместо 90 млн. во втором и 60 млн. в первом релизах соответственно), некоторыми функциональными расширениями, а также добавлением возможности просмотра статистических данных, связанных с употреблением и распределением определенной словоформы в корпусе.

Интернет-корпуса (www.sd-editions.com/LALT/home.html, Лейден и http://titus.uni-frankfurt.de/indexe.htm, Франкфурт-на-Майне) и электронные библиотеки (www.digilib.am, Ереван) существуют для древнеармянского языка; Анаид Донабедян (INALCO, Париж) занимается разработкой корпуса западноармянского языка [Donabédian, Boyacioglu 2007]. Кроме того, имеются электронные библиотеки на современном восточноармянском языке (например, www.armenianhouse.org, www.hayeren.hayastan.com, www.artgrak.am, www.people.cornell.edu, www.arlis.am, www.brusov.am и т.д.). Все эти ресурсы использовались при создании ВАНК; они составляют около 5% художественных и нехудожественных текстов корпуса. Значительный объем восточноармянских публицистических текстов содержится в архивах интернет-изданий www.azg.am, www.aravot.am, www.yerkir.am, www.iravunk.com и т. д.; эти ресурсы легли в основу подкорпуса современной прессы. Попыток создания электронных корпусов восточноармянского языка ранее не предпринималось.

1. Состав корпуса ВАНК

В ВАНК вошли тексты разных жанров, в том числе проза, поэзия, официальные, научные, религиозные, публицистические тексты, а также устная речь современного Еревана. Функциональность выбора подкорпуса (см. ниже) позволяет ограничивать поиск отдельными жанрами и группами жанров, которые интересуют пользователя в данный момент. В целом, ВАНК проектировался таким образом, чтобы максимально полно отразить лингвистическое разнообразие современного восточноармянского языка (см. Приложение 1. Состав ВАНК (на март 2009 г.)).

Важнейшим аспектом корпусной лингвистики являются микроисторические исследования, ориентированные на «быстрые» языковые изменения. Для таких исследований корпус должен иметь временную координату, по которой можно отслеживать изменения значений лексемы или граммемы, отмирание старых и появление новых конструкций. ВАНК покрывает весь новый период истории армянского письменного языка с самого начала аш-

харабара (Хачатур Абовян «Раны Армении» 1841 г.). Временные характеристики текстов также можно использовать при выборе подкорпуса — например, работать только с текстами двадцатого или второй половины двадцатого века.

В идеале каждый жанр должен быть относительно равномерно распределен по годам — или, если это не так, существующая неравномерность должна отражать культурную ситуацию данного периода. В качестве примера препятствующих такой временной сбалансированности технических ограничений можно привести распределение прессы ВАНК по годам. Значительная часть прессы относится к постсоветскому периоду за счет широкой представленности текстов открытых периодических интернет-изданий (около 35 млн. словоупотреблений), доступный объем которых практически неограничен. Эта проблема была отчасти преодолена во втором релизе после осуществления совместного с Национальной библиотекой Армении проекта, в рамках которого были отсканированы, распознаны и включены в корпус избранные выпуски 60 периодических изданий общим объемом более 12 млн. словоупотреблений. Таким образом, архив периодики ВАНК покрывает всю историю существования армянской прессы, начиная от 70-х годов 19-го века по поздний советский период. Тем не менее, временной баланс прессы остается неидеальным.

Важнейшим жанром письменных текстов является электронная коммуникация: электронная почта, смс, instant messengers, блоги. Тексты этого типа в ВАНК только начинают подключаться: в третьем релизе будет содержаться небольшой корпус блогов. Здесь основное затруднение заключается в том, что до самого последнего времени в текстах такого типа использовались в основном разного рода косвенные и мало стандартизованные способы передачи армянской письменности.

Специального обсуждения заслуживает проблема устного корпуса. Иногда возникает вопрос, зачем устные тексты вообще были включены в ВАНК. Претензии, которые предъявляются к устному корпусу (не только в ВАНК, но и, например, Национальному корпусу русского языка) — это отклонение от языковой нормы, массовое использование английских и русских слов, нарушения корректных синтаксических структур. Те же доводы часто приводятся против включения в корпус текстов электронной коммуникации. Все эти претензии, на самом деле, апеллируют не к недостаткам, а к языковым особенностям разговорной речи, которая имеет собственную, иногда значительно отличную от литературной норму, широко использует переключение кода, обладает собственным синтаксисом и т.п. Именно для исследования этих особенностей устной речи и формируются подобные корпуса. Эти исследования относятся не только к лингвистике, но и к смежным областям — социолингвистике (переключение кода), психолингвистике (особенности построения высказывания). Нарушения литературной нормы, особенно если они носят системный характер, должны использоваться при языковом планировании и разработке языковых реформ: такие реформы, которые противонаправлены вектору развития устной речи, обречены на провал. Иными словами, устная речь не является неправильной, не нормативной письменной — она просто другая, иная языковая субстанция.

Отсюда вытекает ответ и на другой вопрос, связанный с сбалансированностью корпуса — как определить правильное соотношение между количеством письменных и устных текстов? Теперь ясно, что это вопрос бессодержательный. Письменный и устный подкорпуса могут находится в произвольном количественном отношении между собой, так как по сути это два разных способа существования языка, два разных корпуса, одновременный поиск по которым имеет ограниченную научную ценность.

С весны 2008 г. на сайте открыт раздел электронной библиотеки, содержащий полные тексты более 100 произведений классической армянской литературы. От других электронных библиотек библиотека ВАНК отличается наличием лексико-морфологического анализа словоформы для всех разбираемых словоформ (более 90% словоупотреблений), для большей части из которых даются также английские переводные эквиваленты (более 85% словоупотреблений).

2. Грамматический словарь

В основу грамматического словаря ВАНК положен словник объемом около 80 тыс. слов. Этот словник является компиляцией многих источников — в первую очередь словаря Е. Г. Галстян [Галстян 1985] и части словаря Э. Б. Агаяна [Агаян 1976], но также словаря аббревиатур Д. С. Гюрджиняна и Н. А. Экекян [Гюрджинян, Экекян 2007], словаря географических названий А. Гргеаряна и Н. Арутюнян [Гргеарян, Арутюнян 1987–1989], различных имен собственных и пр.

Составление грамматического словаря — трудо- и времяемкая работа, которая ранее на армянском материале, насколько нам известно, не проводилась; проект, который решал небольшую часть этой задачи — словарь форм множественного числа [Гюрджинян 2005]. Для сравнения скажем, что первый и достаточно полный грамматический словарь русского языка, содержащий свыше ста тысяч лексем, появился уже более тридцати лет назад [Зализняк 1977].

Несмотря на богатую традицию грамматического описания армянского языка, в готовом виде получить из какого-либо источника классификацию

именных или глагольных парадигм, пригодную для автоматического анализа текстов, невозможно. Для построения алгоритма лемматизации армянских именных словоформ мы выделили более 50 формальных типов именного словоизменения. Полный список *всех* различных парадигматических типов приведен на сайте проекта в разделе «Разметка».

3. Поисковая функциональность ВАНК

ВАНК представляет гибкую функциональность для лингвистического поиска, ориентированную в первую очередь на лексические и грамматические запросы. Синтаксические запросы возможны лишь опосредовано, так как корпус не имеет синтаксической разметки. По поисковой функциональности ВАНК очень близок Национальному корпусу русского языка, который до определенной степени служил его прототипом.

1. ***Поиск словоформы или лексемы.*** ВАНК позволяет искать как вхождения конкретной словоформы (например, մարդու *mardu человек.gen*), так и вхождения всех словоформ определенной лексемы (например, словоформ մարդ *mard человек.nom*, մարդու *mardu человек.gen*, մարդիկ *mardik человек.pl.nom* и т.д. от лексемы մարդ *mard человек.nom*).
2. Вхождения лексем можно искать по их ***английским переводным эквивалентам.***
3. ***Поиск по грамматическим признакам***. ВАНК позволяет искать все словоформы, обладающие определенной грамматической характеристикой или набором грамматических характеристик (например, имперфективный конверб в пассиве). Грамматические признаки можно искать как вне зависимости от того, в какой лексеме они встретились, так и вместе с лексемой. При поиске можно учитывать словоизменительный тип словоформы. Грамматический запрос может:
    a. быть определен как логическая конъюнкция или дизъюнкция нескольких категорий, или
    d. совмещать конъюнкцию и дизъюнкцию в одной логической формуле; собственно, именно последний тип грамматического запроса является наиболее частотным и естественным.

Кроме этих, собственно лингвистических параметров поиска, можно использовать дополнительные графематические и иные параметры, иногда позволяющие эффективно сузить запрос. Так, можно искать только словоупотребления в начале или в конце предложения, накладывать определенные ограничения на регистр (написание с первой прописной или со всеми прописными), указывать знаки препинания слева и справа от вхождения и т.п. ВАНК позволяет искать контексты, в которые одновременно входит несколько поисковых элементов. Расстояние между вхождениями можно изменять, меняя интервал допустимых расстояний.

Сравнивая поисковую функциональность ВАНК с поисковой функциональностью его ближайшего аналога, Русского национального корпуса, можно отметить следующие отличия. ВАНК менее гибок в смысле отрицания словоформ и лексем, но зато представляет возможности отрицания граммем. Он также представляет более гибкие механизмы поиска с учетом регистра и пунктуации, позволяет искать вхождения, находящиеся в начале, в конце или не в начале и не в конце предложения, а также только такие вхождения, которые не имеют омонимичных разборов, причем в ВАНК 3.0 различается внутрилексемная (грамматическая) и межлексемная (лексическая) омонимии. Отсев вхождений с омонимичными разборами в некоторых случаях позволяет сократить количество поискового шума.

Любой запрос, который может быть применен к ВАНК, может быть также применен и определенному пользователем подкорпусу ВАНК. Окно подкорпуса состоит из следующих трех основных зон: авторы и произведения, период, жанр, и трех дополнительных: проза/поэзия, оригинальные / переводные тексты, детская / общая литература.

4. Отображение результатов ВАНК

ВАНК позволяет осуществлять сортировку контекстов по целому ряду параметров: начальная форма словоформы-вхождения (лексема), словоформа-вхождение, словоформа слева от словоформы-вхождения, автор, название, год создания (как по возрастанию, так и по убыванию), жанр. При этом ВАНК поддерживает четыре формата отображения найденной информации:

1. *полный* (по умолчанию): каждый контекст сопровождается базовыми библиографическими сведениями (автор, название, год создания);
2. *краткий*: библиографические сведения приводятся только в окне расширенного контекста;
3. *KWIC (Key Words In Context)*: принятый в корпусных интернет-ресурсах способ отображения контекстов таким образом, чтобы они были визуально выровнены друг относительно друга по вхождению. Формат KWIC используется обычно вместе с сортировкой по словоформе или левой словоформе (см. Приложение 2).
4. *глоссированный*: этот формат предназначен в первую очередь для лингвистов-типологов и изучающих армянский язык. Отображение текста близко к так называемому морфологическому глоссированию (interlinear morphological

glossing), используемому в типологических публикациях и описаниях малых языков, но без разбиения на морфемы и поморфемного перевода. Для всех словоформ, за исключением словоформ, которые не разбираются парсером ВАНК, на экран в виде столбца, расположенного непосредственно под лексемой, выводится лексико-грамматический анализ, который в других типах выдачи доступен только при наведении мыши. В первой строчке столбца содержатся исходная форма и лексические признаки (например, частеречная характеристика). Во второй строке в фигурных скобках приводятся грамматические (словоизменительные) признаки словоформы (за исключением неизменяемых лексем). Если лексеме приписан перевод, он дается в третьей строчке. Если у словоформы существует несколько разборов, они отделяются друг от друга светло-серой чертой (см. Приложение 3).

Отображение результатов возможно как в армянском алфавите, так и в транслитерации (см. Приложение 4). Используемая в ВАНК транслитерация в основном следует международной арменоведческой традиции Хюбшманна-Мейе, адаптированной под Unicode. Транслитерация используется в том числе при отображении имен авторов и названий произведений.

Каждый контекст представлен в окне выдачи одним предложением (исключением является поиск, при котором областью поиска является документ); слова-вхождения при этом выделены оранжевым цветом. При каждом контексте приводятся базовые библиографические характеристики (если они известны) — автор, название, год создания, для прессы также номер или дата выпуска. ВАНК позволяет расширить контекст найденного предложения. По умолчанию на экран выводятся три предложения — то предложение, в котором обнаружены искомые вхождения, а также одно предложение до него и одно предложение после него. Расширяя контекст, можно увеличивать размер контекста вплоть до девяти предложений (четыре предложения до и четыре предложения после того предложения, в котором обнаружено вхождение).

При каждом запросе в верхней части экрана отображается общая информация о запросе и полученных результатах (см. Приложение 2):

- число вхождений (в случае контекстного запроса с числом контекстов более 10,000 — примерная оценка их общего числа в корпусе),
- число документов (в случае контекстного запроса с числом контекстов более 10,000 — примерная оценка числа документов, в которых они могут встретиться),
- критерии сортировки (если они выбраны пользователем),
- размер подкорпуса, по которому осуществлялся поиск (в процентах от общего числа словоупотреблений в корпусе).

Кроме того, в третьем релизе ВАНК добавлена новая функциональность — интерфейс, с помощью которого пользователь может получить не только общую информацию о количестве вхождений словоформы в корпусе, но и подробную картину ее распределения по основным жанрам и декадам (см. Приложение 6. Статистическое употребление словоформы աստծո *astco* (*бог.gen)* по данным ВАНК на март 2009 г.). Кроме абсолютного числа употреблений словоформы, на сайте приводятся еще две характеристики: WPM (число вхождений на миллион), которая позволяет получить представление о частотности словоформы с учетом объема корпуса, а также ее ранг (логарифм отношения частотности самой частотной словоформы к частотности данной словоформы), которая показывает, насколько данная словоформа менее частотна, чем самая частая словоформа данного сегмента. Для краткости приводится только таблица значений показателя WPM, значения которого легче всего интерпретировать.

5. Замечания о программном обеспечении ВАНК

Программное обеспечение для проекта ВАНК разрабатывается и поддерживается компанией Corpus Technologies. Оно создавалось с учетом перспективы масштабирования корпуса, а в конечном итоге — с целью создания языково-независимой программной платформы для корпусных исследований.

Система спроектирована таким образом, чтобы сделать возможным индексирование корпусов разноструктурных языков; проиндексированный системой корпус обеспечивает эффективную обработку запросов разной степени сложности. Единственным, но необходимым требованием является следование разработанному Corpus Technologies стандарту разметки текстов. Только парсер ВАНК и пользовательский интерфейс жестко ориентированы на армянскую грамматику, все остальные структурные элементы системы могут работать практически с любым морфологическим типом языка и алфавитом и разметками разной степени детальности и глубины.

Отметим также алгоритм рандомизации, реализованный далеко не во всех современных крупных корпусах. На настоящий момент пользовательская выдача имеет ограничение 10 тыс. контекстов на запрос. Если число удовлетворяющих запросу контекстов превышает этот лимит (например, при поиске частотной лексемы или отдельного грамматического признака), поисковая система ВАНК использует специальную процедуру, позволяющую избежать нежелательной «конденсации» найденных контекстов в определенной части корпуса и работать с квазирепрезентативной выборкой примеров, более или менее равномерно покрывающей весь корпус.

6. Целевая аудитория ВАНК

Аудиторией проекта является в первую очередь сообщество арменистов, работающих с лексикой и грамматикой ашхарабара, а также специалисты по западноармянскому языку или грабару, исследования которых носит сравнительный характер. С точки зрения представленности различных форм и временных срезов, ВАНК покрывает значительную часть языкового материала и ограничен почти только рамками логически невозможных (несовременные устные тексты) или крайне труднодоступных (жанр частной переписки) типов текстов. Важно подчеркнуть, что корпусом могут пользоваться исследователи, не владеющие или не вполне владеющие армянским языком — арменисты, которые только начинают изучать армянский язык, а также лингвисты-типологи, вообще не специализирующиеся в армянской филологии. Кроме возможности ввода запросов в латинской транслитерации (виртуальная клавиатура) и переключения в латинскую транслитерацию отображения армянской графики, во втором релизе в подсветку грамматического разбора (появляющуюся при наведении мыши на словоформу) включен краткий список английских переводных эквивалентов и псевдоглоссированная выдача (см. Приложение 3).

Благодаря включению в разметку английских переводов пользование корпуса значительно упростилось и для изучающих армянский язык, как лингвистов, так и нелингвистов. Студент-лингвист может проводить собственные микроисследования, пользуясь корпусом точно так же, как и опытный исследователь-арменист, но при необходимости обращаясь к грамматическим разборам и переводам незнакомых форм.

Важной частью целевой аудитории для корпуса, как мы надеемся, могут стать преподаватели армянского языка как иностранного и школьные и вузовские преподаватели армянского языка как родного. Использование корпусов в преподавании — вполне активная, а количественно — чуть ли не доминирующая сфера использования языковых корпусов (см. [Добрушина 2005, 2008]), поскольку корпус позволяет работать с живым языковым материалом и отойти от традиционных методов обучения, опирающихся на закрытый и ограниченный объем признанной литературной классики.

Здесь естественно также упомянуть о той сфере использования корпусов, которая лежит в "серой» зоне между филологами и преподавателями языка — нормативной лингвистике. Как таковая эта отрасль не принадлежит ни к какому направлению академического лингвистического исследования и относится скорее к общественно-политической, чем научной сфере. Еще раз подчеркивая, что корпус ни в коем случае не является образцом нормы, следует отметить, что именно корпус, представляя действительный языковой узус, может и должен становиться основой для работы над нормой. Именно в корпусе видны тенденции языкового развития, изменения узуса, на которые должно ориентироваться языковое планирование. Языковые реформы, которые оторваны от живых языковых процессов, обречены на фиаско, а если таковые реформы принимаются, то в конечном итоге они будут сметены стихией живого языка. В здоровом социуме лингвистический произвол и "вкусовщина» при языковом реформировании невозможны, так как языковой процесс не поддается законодательному регулированию. И здесь важную роль может сыграть как сам ВАНК, фиксирующий языковые сдвиги на протяжении более чем полутора веков, так и устный корпус ВАНК, демонстрирующий живые языковые процессы и обладающий значительным (по сравнению с устными корпусами многих других языков мира) объемом около 3,5 млн. словоупотреблений.

Для представителей других специальностей — историков, социологов, культурологов и др. — корпус может представлять интерес лишь постольку, поскольку они в своих исследованиях обращаются к языковому материалу (что происходит относительно редко). Речь идет о том, как социальные факторы или исторические процессы отражаются в языке, то есть о своего рода исторической социолингвистике. Собственно социолингвисты в основном работают с современным состоянием языка и составляют собственные микрокорпуса, ориентированные на частные задачи. А вот на частные вопросы об узусе того или иного социального значимого концепта, о том, когда он впервые упоминается в письменных текстах, как распространяется, как отмирает, как меняется его наполнение, ВАНК сможет ответить достаточно однозначно. Здесь гибкая грамматическая и контекстная функциональность поиска оказывается излишней (достаточно поиска по лексемам), зато на первый план выступает репрезентативность корпуса и особенно большой объем прессы, для ВАНК — включение во второй релиз значительного архива армянской периодики (около 47 млн. словоупотреблений), покрывающий весь период ее существования.

Наконец, часть потенциальной аудитории корпусов составляют люди, для которых обращение к корпусу вызвано не профессиональной потребностью, а личным интересом к языковому материалу. Языковая рефлексия, рассуждения об узусе, о значении тех или иных слов, как нам кажется, характерны для интеллигентного человека вообще. Поэтому при разработке интерфейса ВАНК мы пытались сделать его функциональность по возможности интуитивной и прозрачной, а разъяснения того, как искать лингвистическую информацию, максимально неспециальными и свободными от лингвистической терминологии. Пользователь корпуса — нелингвист может искать редкие слова, в значении кото-

рых он сомневается, формы, которые кажутся ему неправильными, но которые он встретил в живой речи или в тексте или наоборот, запрещаемые нормой формы, которые кажутся ему естественными или допустимыми.

7. Перспективы проекта

По полноте и репрезентативности литературного языка ВАНК приблизился к некоторому качественному порогу, преодолеть который не только трудно, но и не необходимо. Возможное осмысленное развитие проекта — включение принципиально новых текстов на армянском языке в широком смысле этого слова — литературных западноармянских, диалектных, средне- и древнеармянских текстов.

Технически было бы важно оптимизировать некоторые типы запросов: например, запросы с отрицанием, обработка которых на настоящий момент занимает значительное время. Высокий уровень грамматической омонимии (17% словоупотреблений), как внутри-, так и межлексемной, характерная для армянской грамматики, позволяет говорить о полезности, если не необходимости, работы по снятию омонимии или хотя бы создания подкорпуса со снятой омонимией (ср. корпус с вручную снятой омонимией в НКРЯ).

Создание синтаксической модели и внесение в корпус синтаксической разметки могло бы резко увеличить число академических областей применимости корпуса. Однако такая работа, в первую очередь автоматический синтаксический парсинг, требует огромной теоретической разработки и не может быть осуществлена в обозримом будущем.

На настоящем этапе исследовательская группа ВАНК приступает к корпусно ориентированным исследованиям армянской грамматики, первым из которых стала разработка словаря глаголов, снабженных полной словоизменительной и словообразовательной информацией. В рамках этого же направления компания CorpusTechnologies, технически и финансово поддерживавшая разработку ВАНК, открыла программу корпусных исследований в арменистике, подробное описание которой размещено на сайте корпуса.

Литература

1. *Ագայան Է. Բ.* Արդի հայերենի բացատրական բառարան [Толковый словарь современного армянского языка]. Т. 1–2. Ереван: 1976.
2. *Галстян Е. Г.* (ред.). Հայ-ռուսերեն բառարան [Армяно-русский словарь]. Ереван: 1985.
3. *Գրգեարյան Ա. Կ., Արտյունյան Ն. Մ.* Աշխարհագրական անունների բառարան [Словарь географических названий]. Ереван: 1987–1989.
4. *Гюрджинян Д. С.* Անուն խոսքի մասերի թվի կարգը արդի հայերենում. Քերականական բառարան-տեղեկատու [Категория числа имен в современном армянском. Словарь-справочник]. Ереван: 2005.
5. *Гюрджинян Д. С., Экекян Н. А.* Հայերենում գործածվող տառային հապավումների բառարան [Словарь. Инициальные аббревиатуры в армянском языке]. Ереван: 2007.
6. *Добрушина Н. Р.* Как использовать Национальный корпус русского языка в образовании? // Национальный корпус русского языка: 2003 — 2005. Результаты и перспективы. М.: 2005. С. 308–330.
7. *Добрушина Н. Р.* (ред.) Национальный корпус русского языка и проблемы гуманитарного образования. Теис: 2007.
8. *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение. М. 1977.
9. *Хуршудян В. Г., Подлесская В. И.* Армянское ban как дискурсивный маркер речевого сбоя // Армянский гуманитарный вестник № 1. Ереван: 2006. С. 21–42.
10. *Хуршудян В. Г.* Средства выражения хезитации в устном армянском дискурсе в типологической перспективе. Дис. … канд. филол. наук. М.: РГГУ. 2006.
11. *Donabédian, A., Boyacioglu, N.* «La lemmatisation de l'arménien occidental avec Nooj» // S. Koeva, D. Maurel, M. Silberztein (eds.) Formaliser les langues avec l'ordinateur, de INTEX à NooJ. Presses Universitaires de Franche-Comté. 2007. P. 55–75.
12. *www.aravot.am* — ежедневная газета, Ереван.
13. *www.arlis.am* — Информационная система армянского законодательства (Armenian Legal Information System (ARLIS).
14. *www.armenianhouse.org* — электронная библиотека.
15. *www.artgrak.am* — иностранная литература на армянском языке.
16. *www.azg.am* — ежедневная газета, Ереван.
17. *www.brusov.am* — Ереванский гос. лингвистический университет им В. Брюсова.
18. *www.digilib.am* — электронная библиотека древнеармянских текстов.
19. *www.eanc.net* — Восточноармянский национальный корпус.

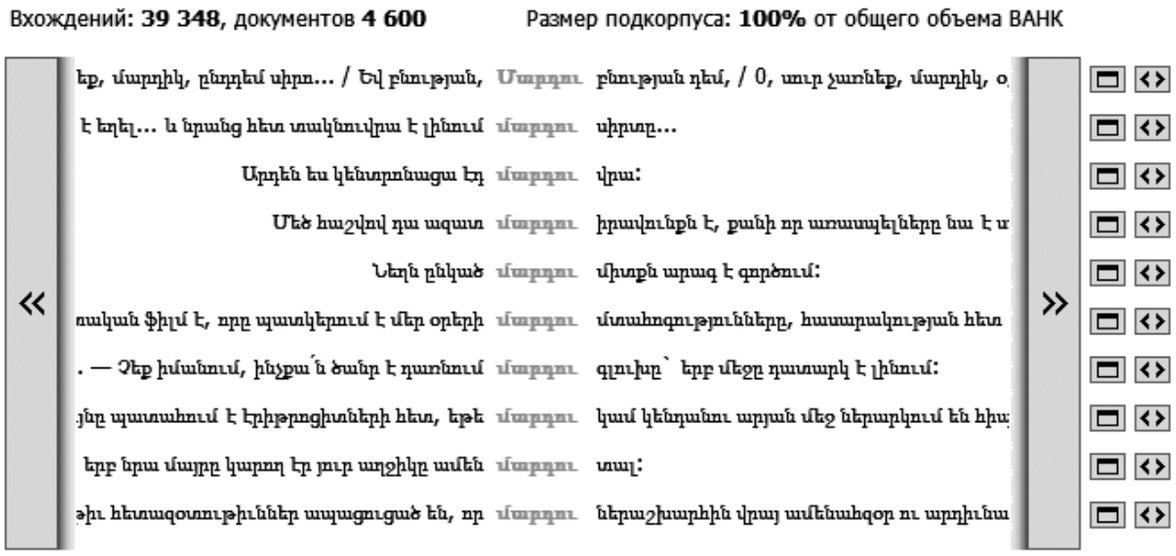
20. *http://forum.am-kayq.com* — армянский форум.
21. *www.hayeren.hayastan.com* — армянский образовательный портал.
22. *www.iravunk.com* — ежедневная газета, Ереван.
23. *www.people.cornell.edu* — текст Библии на восточноармянском.
24. *www.ruscorpora.ru* — Национальный корпус русского языка.
25. *www.sd-editions.com/LALT/home.html* — Leiden Armenian Lexical Textbase.
26. *http://titus.uni-frankfurt.de/indexe.htm* — Thesaurus Indogermanischer Text- und Sprachmaterialien.
27. *www.yerkir.am* — еженедельная газета, Ереван.

Приложение 1. Состав ВАНК (на март 2009 г.)

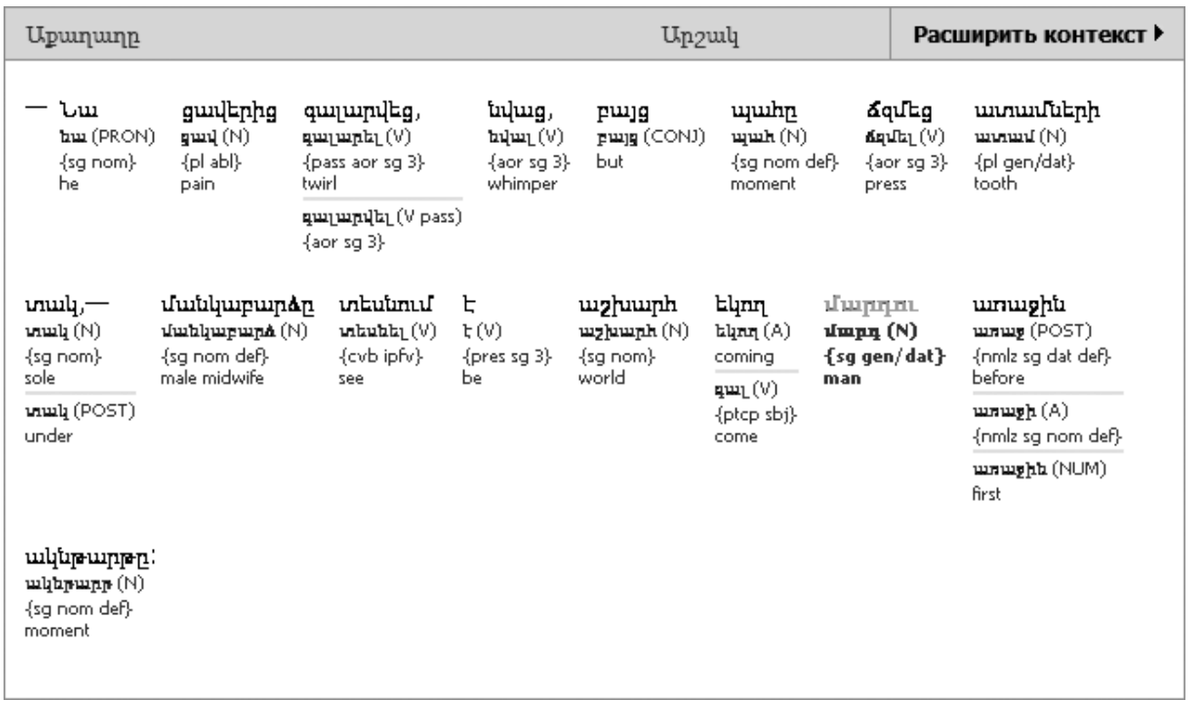
	Письменные тексты	словоупотребления	доля в ВАНК	документы	
	Художественная литература				
	проза: романы	29,909,172	27.1%	371	вкл. 99 переводных
	проза: рассказы	5,959,142	5.4%	183	вкл. 56 переводных
	проза: драматургия	1,411,030	1.3%	55	вкл. 8 переводных
	итого прозы	**37,279,344**	**33.8%**	**609**	
	поэзия	3,648,160	3.3%	227	вкл. 43 переводных
	Пресса	47,264,735	43.0%	7858	
	Нехудожественные тексты				
	научные тексты	13,875,930	12.6%	113	вкл. 22 переводных
	эссе, мемуары, официальные и религиозные тексты	4,735,997	4.3%	379	вкл. 8 переводных
	Итого письменных текстов	**106,804,166**	**96.8%**	**9,186**	

	Устная речь	словоупотребления	доля в ВАНК	документы
	Спонтанная устная речь	1,029,646	0.94%	208
	Публичная устная речь	1,933,899	1.76%	543
	Стимулированные нарративы	70,010	0.06%	22
+	Электронная коммуникация	442,399	0.40%	1
	Итого устной речи	**3,475,954**	**3.2%**	**774**
	Итого в ВАНК	**110,280,120**	**100%**	**9,960**

Приложение 2. KWIC выдача



Приложение 3. Глоссированная выдача



Приложение 4. Отображение результатов в транслитерации

Gorc, 1990.08 #21	1990	**Расширить контекст**
«Tariner aŕaǰ,– asel ē na,– Erewanum Ēdvard Harutʻyunyani glxavorutʻyamb, orn ayžm kendani čʻē, himnel enkʻ **mardu** iravunkʻneri paštpanutʻyan hanjnaxumb:		

Mǰnašen	Xalapʻyan Zorayr	**Расширить контекст**
— Miangamayn aŕoǧǰ **mardu** , očʻ mi hogekan šeǧum:		

Azg, 12.20	2006	**Расширить контекст**
Manuk Gasparyan-Xozi misə hamov ē, baycʻ nra ararkʻnerə, gorcoǧutʻyunnerə, ōrinakʻ mštapes cʻexi meǰ tʻavalvelə, thač en **mardu** hamar:		

Приложение 5. Статистика запроса

Вхождений: **46 477**, документов **5 700** Размер подкорпуса: **100%** от общего объема ВАНК

Մետաքսի ճանապարհը	Շեկոյան Արմեն	**Расширить контекст**
Արդեն չորս տարի է՝ Կիրակոսը Կոմայգի է գալիս ամեն օր, ռեժիմով, ինչպես, ասենք, **մարդիկ** աշխատանքի են գնում:		

Ջերմանց միխիթարություն	Խալափյան Զորայր	**Расширить контекст**
Սկսել եմ ձիու քայլերով ման գալ, **մարդիկ** էլ այքիս շահմատի քարեր են երևում:		

Նրա ճանապարհը, մաս 4	Թաթիկյան Շահեն	**Расширить контекст**
Վատ **մարդիկ** չեն, կծանոթացնեմ:		

Приложение 6. Статистическое употребление словоформы աստծո **astco (бог.gen) по данным ВАНК на март 2009 г.**

Словоформа: աստծո Число вхождений: **9,229** Ранг: **386.3** WPM: **550.**

	Худ.	Нехуд.	Пресса	Устные	Всего по декаде
(1800–1859)	2	n/a	0	n/a	2
(1860–1869)	17	n/a	0	n/a	17
(1870–1879)	61	10	0	n/a	71
(1880–1889)	129	6	0	n/a	135
(1890–1899)	57	n/a	n/a	n/a	57
(1900–1909)	46	10	0	n/a	56
(1910–1919)	92	n/a	0	n/a	92
(1920–1929)	15	0	2	n/a	17
(1930–1939)	167	2	6	n/a	175
(1940–1949)	82	12	0	n/a	94
(1950–1959)	205	18	11	4	238
(1960–1969)	1160	19	30	11	1220
(1970–1979)	759	198	11	2	970
(1980–1989)	906	143	21	6	1076
(1990–1999)	237	77	59	1	374
(2000–2009)	166	528	1673	201	2568
недатированные	2038	15	1	13	2067
Всего по жанру	**6139**	**1038**	**1814**	**238**	9229