*М. Даниэль, Д. Левонян, В. Плунгян,*
*А. Поляков, С. Рубаков, В. Хуршудян*

# ВОСТОЧНОАРМЯНСКИЙ НАЦИОНАЛЬНЫЙ КОРПУС
*www.eanc.net*

M. Daniel, D. Levonian, V. Plungian, A. Polyakov, S. Rubakov, V. Khurshudian
EASTERN ARMENIAN NATIONAL CORPUS

*Eastern Armenian National Corpus is a comprehensive linguistic database of texts in Standard Eastern Armenian. EANC is an annotated corpus of about 110 mln tokens provided with a powerful search engine for making complex lexical morphological queries. The corpus is primarily intended for linguists and Armenian language teachers but can also be used in typological, anthropological and historical studies as well as by anyone interested in Eastern Armenian and its history.*

Восточноармянский национальный корпус (ВАНК) - это открытый интернет-ресурс на базе представительной совокупности текстов на современном восточноармянском языке, предлагающий гибкую функциональность для поиска примеров различных языковых форм. ВАНК, включающий около 110 млн. словоупотреблений, предназначен, в первую очередь, для использования лингвистами-арменистами, во вторую - преподавателями армянского языка, но может представлять ценность и для лингвистов-типологов, историков, культурологов, публицистов и всех тех, кто работает с армянским языковым материалом или интересуется армянским языком и его историей.

**Что такое корпус и чем он отличается от электронной библиотеки**

Первый вопрос, на который следует ответить - что такое электронный корпус языка и чем он отличается от электронных библиотек (таких, например, как доступные на популярных сайтах www.armenianhouse.org или www.hayeren.hayastan.com). На первый взгляд, и корпус, и электронная библиотека представляют собой просто электронную коллекцию текстов. Различие между этими типами ресурсов целиком определяется их предназначением. Корпус предназначен для поиска языкового материала, а библиотека - для чтения текстов. Все остальные различия, определяющие функциональность, интерфейс и другие аспекты реализации этих двух типов электронных ресурсов, являются логическими следствиями различия в их предназначении.

Библиотека должна предоставить пользователю электронные версии текстов в полном объеме. Главная задача разработчиков библиотеки - удобство отображения текстов на экране и навигации, т.е. перемещения по фрагментам текстов и между текстами (система оглавлений, перелистывания и т.п.). Основная проблема,

стоящая перед разработчиками электронных библиотек - это авторские права на размещаемые тексты. Некоторые русскоязычные ресурсы библиотечного профиля прекратили свое существование именно вследствие того, что размещение в интернете текстов в полном виде нарушало права интеллектуальной собственности.

Перед разработчиками корпуса проблема авторского права не стоит, так как они предоставляют своим пользователям лишь краткие фрагменты текстов - обычно одно предложение, иногда, как в случае ВАНК, с возможностью просмотра непосредственного контекста (несколько предложений вокруг найденного). Непосредственный контекст бывает необходим для изучения некоторых языковых явлений: например, при исследовании функций анафорических местоимений, кореферентного опущения и т.п. С другой стороны, корпус должен удовлетворять определенным требованиям, которые ставят достаточно сложные задачи (а) методики сбора материала, (б) его лингвистической обработки и (в) эффективной реализации поисковых алгоритмов.

*Методика сбора материала*. Корпус призван представить язык во всей его полноте, поэтому он должен покрыть как можно большее разнообразие типов текстов и текстовых жанров. Именно поэтому разработчики Британского национального корпуса, одного из первых электронных корпусов, приняли в свое время решение о том, что крупные литературные произведения будут представлены в корпусе лишь фрагментами. В противном случае для английского языка с его огромным литературным наследством либо классические литературные тексты вытеснили бы на периферию другие жанры, либо объем корпуса вышел бы за рамки алгоритмически допустимого (на то время). Ясно, что такого рода решения для электронных библиотек недопустимы. Кроме того, электронная библиотека не обязана представлять тексты второстепенных и третьестепенных авторов, а также тексты документальных или иных "скучных" жанров (от официальных документов до медицинских предписаний), а для корпуса включение таких текстов, по крайней мере в некотором объеме, желательно, так как они представляют собой одну из форм существования письменного языка. Специализированные корпуса (например, корпус газетных текстов) по определению ограничены одним жанром или типом текстов, но в рамках этого жанра они все равно стремятся к максимально полному охвату языкового материала (от аналитических изданий до желтых листков); электронная же библиотека может представить полное собрание сочинений одного крупного автора, но игнорировать другого, менее популярного, по выбору составителей библиотеки и в соответствии с интересами ее аудитории. Иными словами, состав библиотеки определяется тем, что интересно читателю, а состав корпуса - стремлением к возможно более полному покрытию языкового разнообразия.

Существует несколько принятых, но не вполне формализованных характеристик, описывающих электронную коллекцию. текстов, на которую опирается корпус. *Репрезентативным* называется корпус, представляющий языковые выражения (словоформы, лексемы, грамматические категории, словосочетания) во всем разнообразии их узуса - например, в разных жанрах

письменных текстов. *Сбалансированным* называется корпус, в котором соотношение между объемом текстов различных жанров не является произвольным, а опирается на какие-то содержательные критерии. *Полным* называется корпус, который представляет большую или значительную часть существующих письменных текстов данного языка. ВАНК является первым и пока единственным репрезентативным, сбалансированным и полным электронным корпусом современного восточноармянского языка. Такие корпуса иногда называются "национальными" - в том смысле, что они максимально полно представляют языковое наследие данной письменной традиции (Британский национальный корпус, Чешский национальный корпус, Национальный корпус русского языка и др.).

Строго говоря, ВАНК является первым и единственным электронным корпусом современного восточноармянского языка вообще. В сети размещено лишь несколько интернет-корпусов древнеармянского языка (это, например, открытые сайты проектов www.sd-editions.com/LALT/home.html и http://titus.uni-frankfurt.de/indexe.htm).

*Лингвистическая обработка*. Будучи в первую очередь инструментом лингвистического исследования, корпуса требуют достаточно глубокого лингвистического анализа языка. Простейший поиск - поиск слов в том виде, в котором они встречаются в тексте или поиск фрагментов словоформ - обеспечивается любым текстовым редактором и не требует специализированного корпусного обеспечения. Но функциональность корпуса шире, он ищет не только определенные словоформы, но и лексемы в любой форме (*գնալ* ***gnal*** 'идти'), и одинаковые грамматические формы разных лексем (например, императивы множественного числа от непереходных глаголов) - см. в приложении описание функциональности ВАНК. Для этого необходимо программное обеспечение, которое умеет осуществлять грамматический анализ словоформ - так называемые лемматизаторы, которые словоформе W сопоставляют исходную форму лексемы L (это позволяет находить словоформу W по запросу на поиск лексемы L) и ее грамматические и лексические характеристики G1, G2 ... Gn, L1, L2 ... Ln (это позволяет находить словоформу W по запросу на поиск грамматической или лексической категории или их сочетания). Создание программы-лемматизатора является сложной задачей не только для плохо описанных языков, но и для языков с богатой грамматической традицией. Подробнее об этом смотри ниже в разделе о грамматическом словнике ВАНК.

*Эффективные поисковые алгоритмы*. Корпус должен обрабатывать запрос и предоставлять пользователю интересующие его примеры в разумное время (время отклика). Большинство лингвистически содержательных запросов обрабатываются ВАНК в течение нескольких секунд. Время отклика является техническим требованием к ресурсу. То, насколько сложные алгоритмы потребуются для удовлетворения этого требования, зависит, во-первых, от объема корпуса и, во-вторых, от степени гибкости предполагаемой поисковой функциональности.

Реализация поиска на многомиллионных корпусах с функциональностью, аналогичной функциональности ВАНК, является очень сложной алгоритмической задачей, сопоставимой по сложности (и отчасти аналогичной) поисковым интернет-системам типа Google.

Как видно из вышесказанного, составители электронных библиотек и составители корпусов сталкиваются с проблемами разного рода; причем последние должны решать большее число научных, методологических и алгоритмических задач. Конечно, и тем, и другим на начальном этапе приходится осуществлять одну и ту же, технически сложную и затратную работу по оцифровке бумажных версий текстов - сканирование и автоматическое распознавание либо набор с последующей корректурой. В случае ВАНК было оцифровано более 70 млн. словоупотреблений; остальной объем (чуть менее 40 млн., из которых около 35 млн. составляет электронная пресса) был загружен с открытых интернет-сайтов.

В связи с обсуждением электронных библиотек отметим, что на сайте ВАНК открыт специальный раздел - Электронная библиотека ВАНК, в которой в полнотекстовом доступе размещено более ста произведений классической армянской литературы. Все размещенные в библиотеке произведения были оцифрованы в рамках проекта, поэтому значительных пересечений с другими общедоступными ресурсами (уже упомянутыми выше www.armenianhouse.org и пр.) нет. От других электронных библиотек библиотека ВАНК отличается наличием лексической и морфологической разметки, которой может пользоваться, например, студент, изучающий армянский язык. Если щелкнуть мышкой по словоформе, на экране появляется исходная форма лексемы, характеризующие данную словоформу словоизменительные признаки, для большинства словоформ английские переводы, а также некоторая другая лингвистическая информация.

**Состав корпуса ВАНК**

Как уже было сказано, чтобы быть эффективным инструментом лингвистических исследований, корпус должен представлять языковое выражение во всем спектре его употреблений и контекстов и отражать относительную частотность этих употреблений, то есть быть репрезентативным и сбалансированным. Несмотря на то, что языковая репрезентативность корпуса - характеристика трудноформализуемая, ясно, что корпус должен по крайней мере представлять жанровое разнообразие языка и что наиболее важные жанры - проза, поэзия, публицистика - должны быть представлены в нем достаточно полно. Корпус, который содержит только публицистические, только научные или только художественные тексты, или корпус, в котором все эти жанры есть, но публицистики, скажем, очень мало, не может полностью удовлетворить лингвиста. Конечно, полнота представленности - понятие также относительное, и поэтических текстов в любом национальном корпусе всегда мало (в ВАНК - менее четырех миллионов словоупотреблений, или около 3%). В ВАНК вошли тексты разных

жанров, не только проза, поэзия и пресса, но и официальные, научные, религиозные тексты. При этом корпус позволяет ограничивать поиск отдельными жанрами и группами жанров, которые интересуют пользователя в данный момент.

Таблица 1. Распределение письменных текстов по основным жанрам

| **Письменные тексты** | **словоупотребления** | **% от ВАНК** |
|---|---|---|
| пресса | 47 265 000 | 43,0% |
| проза | 37 029 000 | 33,7% |
| наука | 13 750 000 | 12,5% |
| другие нехудожественные | 4 681 000 | 4,3% |
| поэзия | 3 627 000 | 3,3% |
| **Всего письменных текстов** | **106 352 000** | **96,8%** |

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 армянской литературы с самого начала ашхарабара в том состоянии языка, в котором он представлен, например, в произведениях Абовяна. Временные характеристики текстов также можно использовать при выборе подкорпуса - например, работать только с текстами двадцатого или второй половины двадцатого века.

Отправляя запрос и получая ответ в виде совокупности контекстов, пользователь ставит себя в зависимость от того, какие жанры и периоды больше, а какие меньше представлены в корпусе. Если, например, научная литература представлена в корпусе непропорционально широко, у исследователя может сложиться неверное представление о частотных характеристиках того или иного языкового феномена. Принято говорить о необходимости баланса разных жанров в составе корпуса. Установленных универсальных пропорций между разными жанрами в корпусной лингвистике не существует. Более того, кажется, что их и не может существовать - культуры могут отличаться по тому, какую роль в их письменном языке играет художественная литература, а какую публицистика, какую проза, а какую поэзия, и т.д. Чаще всего к проблеме баланса подходят именно с такой "культурологической" точки зрения; методов формальной оценки лингвистической сбалансированности корпуса пока, насколько нам известно, не существует. Культурологическая же оценка обычно бывает вполне приблизительна - считается, что пресса и художественная литература должны быть представлены сравнимыми объемами текстов, поэзии вполне может быть заметно меньше, чем

прозы, научной литературы может быть меньше, чем художественной, и так далее. Отметим, что такие пропорции достаточно точно отражают степень доступности текстов того или иного типа.

Однако на пропорциональную представленность в корпусе разных жанров порой накладываются внешние ограничения, в разной степени очевидные и в разной степени неизбежные. Важный тип текстов, почти не представленный в ВАНК - это эпистолярные тексты. Письма писателей и выдающихся людей в некотором объеме в корпусе представлены, а вот переписка простых людей, так сказать, "бытовая" эпистолярика в корпусе отсутствует. Несмотря на то, что доля бытовой письменной корреспонденции в общем объеме написанных на армянском языке письменных текстов значительна, включить такие документы в корпус затруднительно по чисто техническим причинам (трудности с доступом к рукописным документам и с их обработкой).

Интересным и пока слабо охваченным жанром письменных текстов является электронная коммуникация: электронная почта, смс, instant messengers, блоги. Эти тексты представляют интерес, поскольку они представляют новый тип коммуникации, отчасти промежуточный по своим характеристикам между письменным и устным. Основное затруднение заключается в том, что до самого последнего времени в текстах такого типа на армянском языке использовались в основном косвенные и нестандартизованные способы передачи армянской письменности. В марте 2009 г. в корпус был включен первый массив текстов этого рода - армяноязычный форум forum.am-kayq.com.

Но основная сложность заключается в том, что в идеале каждый жанр должен быть относительно равномерно распределен по годам; колебания, если таковые имеются, должны отражать какую-то содержательную характеристику письменной традиции. В качестве примера действия препятствующих такой временной сбалансированности технических ограничений можно привести распределение прессы ВАНК по годам. Значительная часть прессы относится к пост-советскому периоду за счет широкой представленности текстов открытых периодических интернет-изданий (в ВАНК - около 35 млн. словоупотреблений), доступный объем которых практически неограничен, в то время как пресса советского и досоветского периода существует только в виде печатных изданий. Эта проблема была отчасти преодолена 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. Архив периодики относительно равномерно покрывает всю историю существования армянской прессы, начиная от 70-х годов 19-го века по поздний советский период. Но полученная из электронных архивов пресса последних лет все равно дает всплеск, лишь отчасти мотивированный экспансией интернет-периодики. Поэтому баланс прессы по времени (а также соотношение публицистических и иных текстов в последней декаде) остаются не идеальными.

Ср. таблицу:

Таблица 2. Распределение жанров по времени
(объем в тысячах словоупотреблений и проценты от общего числа словоупотреблений в декаде)

| **период** | **проза** | | **поэзия** | | **нехудожественные** | | **пресса** | | **итого за период** |
|---|---|---|---|---|---|---|---|---|---|
| | | **% за период** | | **% за период** | | **% за период** | | **% за период** | |
| до 1870 | 292 | 64% | 4 | 1% | n/a | 0% | 161 | 35% | 456 |
| 1870 - 1879 | 515 | 53% | 49 | 5% | 250 | 26% | 150 | 16% | 963 |
| 1880 - 1889 | 1431 | 74% | 4 | 0% | 48 | 3% | 447 | 23% | 1930 |
| 1890 - 1899 | 802 | 100% | n/a | 0% | n/a | 0% | n/a | 0% | 802 |
| 1900 - 1909 | 736 | 36% | 84 | 4% | 253 | 12% | 955 | 47% | 2029 |
| 1910 - 1919 | 452 | 60% | 62 | 8% | n/a | 0% | 246 | 32% | 759 |
| 1920 - 1929 | 740 | 44% | 297 | 18% | 44 | 3% | 599 | 36% | 1680 |
| 1930 - 1939 | 2211 | 57% | 28 | 1% | 243 | 6% | 1410 | 36% | 3892 |
| 1940 - 1949 | 923 | 46% | 139 | 7% | 199 | 10% | 733 | 37% | 1993 |
| 1950 - 1959 | 2408 | 47% | 785 | 15% | 463 | 9% | 1422 | 28% | 5078 |
| 1960 - 1969 | 4014 | 57% | 479 | 7% | 426 | 6% | 2176 | 31% | 7095 |
| 1970 - 1979 | 5885 | 48% | 122 | 1% | 4355 | 36% | 1899 | 15% | 12262 |
| 1980 - 1989 | 3984 | 34% | 69 | 1% | 5936 | 50% | 1861 | 16% | 11850 |
| 1990 - 1999 | 1227 | 37% | 79 | 2% | 1325 | 40% | 650 | 20% | 3281 |
| 2000 - 2008 | 879 | 2% | 37 | 0% | 3993 | 10% | 34553 | 88% | 39462 |
| недатированные | 10531 | 82% | 1391 | 11% | 897 | 7% | 3 | 0% | 12821 |
| **Итого** | **37029** | **35%** | **3627** | **3%** | **18431** | **17%** | **47265** | **44%** | **106352** |

Для армянской литературной традиции двадцатого века очень важны переводы русской и западной литературы. В корпус включен значительный объем (около 12 млн. словоупотреблений) переводных текстов более 100 авторов, в основном художественная литература, причем ВАНК дает возможность ограничить поиск только переводными (или, соответственно, только оригинальными) текстами. К переводным текстам отнесены, в том числе, переводы на современный армянский

классических древних авторов - Хоренаци, Нарекаци, Даврижеци и некоторых других.

Оригинальная восточноармянская художественная литература представлена произведениями 241 автора и составляет более 35 млн. словоупотреблений (около трети всего корпуса). Для сравнения можно сказать, что объем "Самвела" - менее 150 тыс. словоупотреблений, а полное собрание сочинений Раффи содержит менее полутора миллионов словоупотреблений. В этой связи излишне говорить, что ВАНК покрывает не только все тексты школьной программы, но практически все сколько-нибудь известные произведения армянской литературы. С другой стороны, ВАНК является не собранием критических филологических изданий, а лингвистическим инструментом. Поэтому в корпусе могут отсутствовать периферийные тексты классических авторов и иные случайные лакуны, отсутствуют сноски и комментарии, остается довольно много опечаток.

В целом ВАНК является одним из крупнейших языковых корпусов в мире (ср. краткие описания других корпусов на http://www.linguistlist.org/sp/texts.html).

**Устный корпус ВАНК**

Кроме различных письменных текстов в состав ВАНК входит большой подкорпус устной речи. Устные тексты также диверсифицированы и включают публичную устную речь (в основном расшифровки телепередач - около 1,7 млн. словоупотреблений), спонтанную устную речь (около 1 млн.), а также более мелкие и более специальные жанры - электронная коммуникация (http://forum.am-kayq.com - около 400 тыс. словоупотреблений), речь кино (расшифровка речи актеров - около 200 тыс. словоупотреблений), так называемая стимулированная речь (расшифровки психолингвистических экспериментов - около 70 тыс.), всего почти 3,5 млн. Отметим, что, хотя о балансе и о полноте устного корпуса говорить вообще сложно - множество устных текстов, в отличие от множества письменных текстов, принципиально открыто - основной тип устной речи, спонтанные диалоги, представлены в корпусе достаточно широко. Все устные тексты были записаны и расшифрованы в рамках реализации проекта ВАНК.

Армянский язык, таким образом, вошел в число очень немногих языков, для которых существуют обширные корпуса устной речи; в основном это крупные европейские языки - например, английский, итальянский, русский. Для сравнения скажем также, что, хотя общий объем устных текстов в Национальном корпусе русского языка составляет почти 6 млн., что заметно больше, чем в ВАНК, объем спонтанных устных текстов, которые, с нашей точки зрения, представляют стихию устной речи наиболее непосредственно, в ВАНК более чем вдвое превышает соответствующий подкорпус НКРЯ.

Устная речь - это динамичная и живая субстанция, практически не поддающаяся регулированию или иному целенаправленному воздействию извне и в диахроническом плане опережающая современное ей состояние письменного языка. В ней отражается большое число культурных влияний, социальных изменений, она является котлом из самых разных языковых процессов, а иногда и

разных языков; погружена в прагматику ситуации и потому гораздо менее эксплицитна, более эллиптична; гораздо менее отрефлектирована и потому во всех отношениях менее нормативна, не только и не столько на уровне лексики, сколько на уровне грамматики и особенно синтаксиса. Некоторые ученые считают, что только такая субстанция и является единственным достойным объектом лингвистического исследования, у других устные данные вызывают неприятие. Это верно для лингвистической традиции вообще и для арменистики в частности. Использование разговорных, ненормативных грамматических форм, иностранных, в основном русских или английских слов, жаргонизмов - все это расценивается не просто как отклонение от литературной и письменной нормы, но как "неправильность" языка, который поэтому не является достойным объектом изучения (те же доводы часто приводятся против изучения текстов электронной коммуникации). С первым трудно спорить, но и согласиться со вторым никак нельзя.

Все эти претензии, на самом деле, апеллируют не к недостаткам, а к языковым особенностям устной речи, которая имеет собственную, иногда кардинально отличную от литературной норму, действительно широко использует переключение кода, действительно обладает собственным синтаксисом и т.п. Именно для исследования этих особенностей устной речи и формируются подобные корпуса. Такие исследования обычно относятся не к сфере чистой лингвистики, а к смежным областям - социолингвистике (переключение кода), психолингвистике (особенности построения высказывания, речевые ошибки и т.п.), и в последнее время начинают проводиться на материале всех европейских языков. Примеры исследований такого рода в арменистике см. [Хуршудян 2006; Хуршудян, Подлесская 2006]. Конечно, устная речь значительно отличается от письменной. Но это не значит, что устная речь неправильна - просто она живет по своим собственным законам. Стихия устной речи позволяет изучать тенденции языкового развития. Знание этих тенденций оказывается важным не только для чистой лингвистики и социолингвистики, но и, например, для языкового планирования: осуществление языковой политики в отрыве от живых языковых процессов никогда не бывает вполне успешным. Поэтому мы считаем устный подкорпус важнейшей частью проекта ВАНК и хотели бы привлечь к нему особое внимание исследовательского сообщества.

Отсюда вытекает ответ и на другой вопрос о сбалансированности корпуса - как определить правильное соотношение между количеством письменных и устных текстов? Из вышесказанного ясно, что это вопрос бессодержательный. Объем письменного и устного подкорпусов могут находиться в произвольном отношении между собой, так как по сути это два разных способа существования языка, две грамматики, два корпуса.

**Разметка корпуса**

Как уже было сказано, важнейшее отличие корпуса от любой другой совокупности текстов - наличие грамматической и иной информации, внесенной в

него разработчиками. Такая информация обычно называется *разметкой* (markup). Неразмеченный корпус представляет для лингвиста определенную ценность, но лингвистическая ценность размеченного корпуса, снабженного эффективной системой поиска, на много порядков выше, а корпус, не сопровождаемый лингвистическим аппаратом, в принципе не может обладать эффективной (с точки зрения исследователя-лингвиста) системой поиска.

Технология внесения в корпус лингвистической информации может быть очень различна - от полностью ручной до полностью автоматической. Ручная или частично ручная разметка используется при документации диалектов или малых и вымирающих языков или в специальных проектах с малым объемом текстов, требующих специфической ручной обработки - например, при создании корпусов детской речи и в других тонких психолингвистических исследованиях. В случае национальных корпусов такая технология неприменима просто по причине их большого объема. Тексты ВАНК были обработаны программой лемматизатором, которая сопоставляла каждой словоформе исходную форму и набор лексических и грамматических помет. То, как эта информация записывается, не так важно (в частности, она может храниться и отдельно от самих текстов). Примерное представление о разметке дает следующий пример:

*կայծակը*
**կայծակ (N, inanim)** – *неодуш. сущ., исходная форма: կայծակ*
**{sg nom def}** – *грамматические признаки: единственное число, именительный падеж, определенная форма*
**lightning** – *английский перевод*

Как видно, в записи используются сжатые, условные и отчасти технические обозначения - N вместо существительного, sg вместо единственного числа, nom вместо именительного падежа, def вместо определенной формы: подробнее описание помет см. на сайте в разделе "Разметка ВАНК (список помет ВАНК)". Наличие разметки позволяет поисковому компоненту корпуса быстро находить интересующие пользователя словоформы и показывать контексты, в которых они встретились. Для эффективного и быстрого поиска происходит индексация корпуса, при которой сам корпус преобразуется в базу данных - но все это скрыто от пользователя, принципиальное отличие корпуса от электронной библиотеки заключается именно в наличии лексико-грамматической информации как таковой.

Не все словоформы корпуса разбираются программой-лемматизатором. Во многих случаях это связано с опечатками или ошибками распознавания исходных текстов. Есть и более содержательные ошибки - лемматизатор не в состоянии обработать авторскую орфографическую игру, при которой искажается правильное написание слова; вкрапления из грабара и западноармянского, иностранные слова (переключение кода) и т.п. Наконец, некоторые редкие лексемы могут быть не включены в словник ВАНК, им может быть приписан ошибочный словоизменительный тип или не указан допустимый словоизменительный вариант. На настоящий момент вообще не разбирается 7% словоупотреблений ВАНК; еще

около 3% составляют цифровые вхождения и вхождения использующие цифры других алфавитов. Это не значит, что остальные 90% разобраны правильно - вхождению может быть приписан неправильный разбор - однако такой статистикой мы не располагаем.

Еще одним важным компонентом ВАНК является наличие библиографической разметки, включающей жанр и название текста, время написания или издания, имя автора и некоторые другие сведения. Эта информация важна с двух точек зрения. Во-первых, пользователь корпуса, работая с примерами, должен знать, к какому жанру принадлежит данный текст, а также кем и особенно когда он был написан, так как от этого существенно зависит интерпретация языковых данных. Во-вторых, ВАНК предоставляет пользователю возможность ограничить поиск языкового материала заданным им подкорпусом. Подкорпус можно ограничить разными основаниями - жанр, время написания, автор и т.п. Понятно, что для обеспечения такой функциональности необходимо приписать текстам базовую библиографическую (метатекстовую) разметку. Оговоримся, что сведения о времени написания носят приблизительный характер и нуждаются в доработке - для многих произведений период написания определен лишь очень приблизительно, с точностью до пятидесяти лет. В значительной степени это связано с недоступностью сводных библиографических источников. Жанровая разметка и разметка по авторам произведений проведена достаточно последовательно и хорошо.

**Грамматический словарь**

Человек, читающий текст на иностранном языке, должен открыть словарь, найти возможные исходные формы и принять решение о том, формой какой лексемы является интересующая его словоформа. Примерно так работает и лингвистический аппарат корпуса. Для того чтобы произвести лемматизацию, т.е. внести лексическую и грамматическую разметку, программное обеспечение корпуса должно уметь устанавливать соответствие между словоформой в тексте и исходной формой лексемы. Для этого необходим уже кратко упоминавшийся выше грамматический словарь, в котором приводятся не только исходные формы (леммы) слов, но и их парадигматические типы, исчерпывающим образом определяющие их формальную парадигму.

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.

Однако компиляция словника - это лишь часть работы, и весьма небольшая. Если программа не имеет никакой информации о словоизменении, она не сможет определить, что словоформа *գիտի gini* является именительным падежом от лексемы

*գինի gini* 'вино', а не родительным падежом от лексемы *գին gin* 'цена'. Только наличие в словнике словоизменительных помет позволяет понять, что родительный падеж от *գին gin* образуется с редукцией гласного – *գնի gni.* 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. Дело не в том, что какие-то словоизменительные типы остались неописанными в традиционной арменистике (хотя в ВАНК и встречаются типы, отсутствующие в традиционных грамматиках, но присутствующие в разговорном языке и потому отраженно представленные в литературе - например, *տղի tġi – տղու tġu*). Важнее то, что задачи автоматической обработки текстов имеют приоритеты, отличные от приоритетов теоретического и практического грамматического описания. Для построения алгоритма лемматизации армянских именных словоформ мы выделили более 50 формальных типов именного словоизменения. С лингвистически содержательной точки зрения многие из этих типов могут быть сведены друг к другу или выведены из фонотактики слова. Практически каждому набору падежей соответствует два словоизменительных типа в зависимости от выбора показателя множественного числа *–եր -er/-ներ -ner*, что "несодержательно" удваивает общее число типов. Сам выбор показателя множественности в статистически подавляющем большинстве случаев определяется количеством слогов в исходной форме. Некоторые типы отличаются друг от друга только наличием беглой гласной в корне. Однако для удобства автоматической обработки текстов вместо того, чтобы отягощать морфологическую модель различными содержательными правилами, гораздо проще приписать каждому существительному один из пятидесяти словоизменительных типов в готовом виде. Ясно, что в научных работах задача описания словоизменения таким образом не ставится. Полный список *всех* различных парадигматических типов приведен на сайте проекта в разделе "Разметка".

С практической точки зрения работа над словником была организована по принципу итераций - после того, как на основании различных источников был создан, размечен и выверен исходный вариант словника ВАНК, была проведена первоначальная обработка массива текстов. После этой обработки просматривались наиболее частотные формы, которые лемматизатор разобрать не смог - в основном это были либо словоформы лексем, отсутствовавших в словнике ВАНК, либо последствия неправильно приписанной словоизменительной информации (в

частности, отсутствие факультативного словоизменительного варианта). После этого словник ВАНК поправлялся и исправлялся. Такие итерации время от времени повторяются, после чего лексико-морфологическая разметка корпуса обновляется.

**Морфологическая модель**

В основе любого алгоритма морфологического анализа текста лежит то или иное лингвистическое описание. Необходимо принять решения об интерпретации спорных форм, иногда относительно периферийных, иногда центральных. В ходе исследования лингвист часто сталкивается с неоднозначностью языковых данных (не говоря уже о различиях в теоретических установках) и с возможностью их различной интерпретации; одни и те же формы могут интерпретироваться как словоизменение или словообразование, одни исследователи выделяют один набор частей речи, другие - другой и т.п. При разработке системы автоматического морфологического анализа нужно выбрать одно из возможных решений. Выбор решения прямо или косвенно сказывается на пользователе корпуса - ему приходится адаптироваться под используемую в корпусе модель морфологии и принимать во внимание интерпретации, которые могу казаться ему спорными или неверными по сути.

Примером вопроса о центральных, но спорных категориях, является проблема различения в армянском языке дательного и родительного падежа. Большая часть описаний различает эти формы, однако нельзя не признать, что грамматисты испытывают здесь серьезное давление классической грамматической традиции. Если же обратиться к внутренним языковым свидетельствам, то оказывается, что у имен дательный падеж отличается от родительного только своей способностью присоединять определенный артикль: первый обычно имеет показатель определенности, а второй никогда с ним не сочетается. Не вполне ясно, является ли это достаточным основанием для выделения двух разных падежей. Почти с теми же основаниями в армянском можно было бы выделять, например, специальную счетную форму - употребление формы единственного числа в контексте числительного очевидно имеет здесь особую функцию, которая с трудом может быть выведена из семантики единственности.

Важным доводом в пользу различения падежей могло бы служить то, что в армянском языке есть именная часть речи, у которой форма дательного падежа отличается от формы, которая используется в приименной позиции – это личные местоимения. Однако такие формы личных местоимений можно в принципе считать отдельными лексемами – притяжательными местоимениями. Таким образом, вполне оправдана точка зрения, согласно которой в армянском языке есть единый падеж, выполняющий одновременно функции дательного и винительного падежей; а поскольку он маркирует еще и прямое дополнение, его можно было бы называть просто косвенным падежом, аналогичным обликвусу (oblique), широко представленному во многих языках с бедной падежной системой. Однако в морфологической модели ВАНК принята другая, более традиционная точка зрения, согласно которой притяжательные формы являются элементами падежной

парадигмы местоимений, а у имен выделяются две разные падежные категории, датив и генитив, что не в последнюю очередь обусловлено стремлением сохранить единство структуры парадигмы для всех именных частей речи. Впрочем, лемматизатор ВАНК не может различить родительный и дательный падеж у существительных без артикля (они могут быть различены только в контексте, который лемматизатор не учитывает), поэтому де факто в таких случаях обе грамматические пометы приписываются одновременно: gen/dat.

Примерами периферийных форм, которые встречаются в корпусе, но почти не обсуждаются в традиционных описаниях, являются реляционная форма имени или форма ассоциативной множественности. Реляционная форма имени, или релятив – это относительно редкая форма, более характерная для устной речи, но изредка встречающаяся и в письменных текстах. С морфологической точки зрения релятив – это именная основа, к которой присоединен показатель родительного падежа, затем определенный артикль, а затем либо просто определенный артикль, либо, реже, показатель дательного падежа и определенный артикль или показатель родительного или одного из периферийных (пространственных) падежей.

1. *սեղանինը*
   seġan-i-n-ə
   стол-GEN-DEF-DEF
   ‘то, что на столе’

2. *սեղանիինը*
   seġan-i-n-in-ə
   стол-GEN-DEF-DAT-DEF
   ‘тому, что на столе’

3. *սեղանինով*
   seġan-i-n-ov
   стол-GEN-DEF-INST
   ‘тем, что на столе’

С функциональной точки зрения релятив является субстантивацией формы генитива и может присоединять именные грамматические показатели, например, падеж. Значение такой формы можно описать как 'принадлежащий / имеющий отношение к N', где N – это производящая именная основа. Имеются основания полагать, что артикль, следующий за показателем генитива, с функциональной точки зрения не может считаться артиклем, так как он выступает здесь в роли не артикля, а номинализатора или, в другой терминологии, субстантиватора. Поэтому реляционные формы имени в ВАНК считаются субстантивированными формами генитива.

Формы ассоциативной множественности - это формы, которые обозначают группу лиц, ассоциированную с неким главным членом этой группы, обозначенным производящей именной основой. Форма более характерна для разговорной и диалектной речи. В качестве примеров можно привести, например, *Վարդանանք* ‘Вартан и его группа (его друзья, родственники)’, *Շուշանենք* ‘Шушан и ее группа (ее друзья, родственники)’. Морфологически к формам ассоциативной множественности примыкают местоименные формы вида *մերոնք* ‘наша группа’, хотя с семантической точки зрения их интерпретация как форм ассоциативной множественности несколько проблематична.

Наконец, пользователю могут показаться непривычными или произвольными некоторые решения, относящиеся к сфере грамматической терминологии. Эти решения были направлены в первую очередь на сближение арменистической традиции с современной типологической практикой описания языков. Так, ВАНК

относит к различным частям речи причастия и деепричастия, в названиях падежей используется латинская номенклатура, деепричастие называется более распространенным в теоретической литературе термином конверб и т.п. Пожалуй, самым значимым терминологическим решением является использование вместо термина пассива термина медиопассив. Несмотря на то, что в русской армяноведческой традиции принято говорить о пассиве, на самом деле эта категория в армянском языке гораздо шире - она похожа на русскую возвратность, а в типологической номенклатуре соответствует термину медиальный залог, или медий.

Итак, при работе с ВАНК пользователю отчасти приходится вставать на теоретические позиции лежащей в его основании лингвистической модели. В целом, при известном навыке, это не должно создавать у пользователя серьезных проблем - нужно лишь желание работать с корпусом и готовность принять некоторые "правила игры". Конечно, вопрос не в смене научных взглядов, а только в умении и стремлении пользоваться корпусом как инструментом исследования. Кроме того, речь идет о словоизменительной морфологии и различных лексических пометах, в рамках которой разнообразие интерпретаций значительно более ограничено, чем, например, в сфере синтаксиса. Подробно различные технические решения и конвенции описаны на сайте ВАНК в разделе о разметке.

**О грамматической омонимии и контекстной информации**

Понимание естественного языка - это многокомпонентная деятельность, предполагающая работу сразу многих языковых модулей. Человек, читающий текст, воспринимает слово не вне, а внутри контекста - у него не возникает проблем различить значения одной и той же словоформы *գրում grum* в таких контекстах как:

4. *«Հառա՛ջ, ի փառս հայրենիքի», — հպարտորեն* ***գրում*** *է նա օրագրում:*
*«Вперед, во славу родины!», — гордо пишет он в дневнике.* (С. Цвейг «Звездные часы человечества»)

5. *Գիտնական այր դարձավ, սակայն հոգեւոր ոլորտի հմայքը չկորցրեց նաեւ իր* ***գրում:***
*Он стал ученым, однако в своих произведениях продолжал отдавать должное духовной сфере.* (газета «Азг», 23.01.2004)

Значительная часть лемматизаторов – в том числе лемматизатор ВАНК - устроена иначе. Форма рассматривается в отрыве от контекста, так что словоформе, которая может иметь несколько разных интерпретаций, приписываются все возможные грамматические интерпретации вне зависимости от окружающего контекста. В данном случае в обоих случаях приписывается и разбор словоформы

как локатива от не очень частотного существительного *գիր gir* 'письменность, письменное творчество', и как конверба (деепричастия) несовершенного вида от крайне частотного глагола *գրել grel* 'писать'. При поиске форм, допускающих омонимичные разборы, пользователь должен быть готов к тому, что в результатах нужные ему языковые данные окажутся перемешанными с поисковым "шумом". В приведенном выше случае, если пользователь хочет найти формы местного падежа от лексемы *գիր gir*, шумом окажется подавляющее большинство результатов поиска, так как в значении конверба форма *grum* встречается гораздо чаще, чем в значении падежной формы. Долю шума в результатах поиска можно несколько сократить, используя контекстный поиск. Например, если указать, что перед формой *գրում grum* идет прилагательное, форма родительного падежа или другой приименной атрибут, вероятность нахождения контекстов с локативностью несколько возрастает. Но при таком поиске мы упускаем те контексты, в которых *գրում grum* является формой имени, но не имеет при себе определения.

Аналогично обстоит ситуация и с поиском просто по грамматическим показателям. Если искать формы конверба несовершенного вида от любого глагола, в поиске будут попадаться формы, омонимичные с ними, но являющиеся формами местного падежа (тот же *գրում grum* в значении 'в (своих) книгах'). Как и при поиске словоформ, можно воспользоваться контекстным поиском: ввести такие условия, которые делают одну интерпретацию более вероятной, чем другую. В данном случае можно указать, что непосредственно перед или после искомой формы стоит глагол-связка. Тогда мы исключим заметную долю ненужных употреблений тех форм на *-ում -um*, которые имеют омонимию типа V, cvb ↔ N, loc. Но шум все равно остается, так как форма местного падежа вполне может стоять рядом со связкой, а часть нужных нам контекстов при этом теряется, так как не всегда конверб находится в непосредственном контакте с глаголом-связкой.

Еще один способ сократить шум - выбрать специальную опцию поиска, исключающую из результатов словоформы с неединственным разбором. В этом случае, если мы ищем словоформы лексемы *գիր gir*, мы найдем только именные формы, а если мы ищем словоформы глагола *գրել grel* - только глагольные. Но тогда словоформа *գրում grum* вообще не попадет в результаты поиска, так как у нее есть два разбора - а ведь она составляет значительную часть словоупотреблений глагола *գրել grel*; если исключить вообще все омонимичные разборы, то останется лишь около двадцати процентов всех контекстов с лексемой *գրել grel*.

Наконец, поиск без учета контекста имеет еще одно важное, системное грамматическое последствие. В глагольной парадигме современного армянского языка преобладают аналитические глагольные формы, состоящие из сочетания нефинитной глагольной формы (мы называем ее конвербом) со вспомогательным глаголом (связкой). Однако, поскольку лемматизатор не умеет устанавливать связи между словоформами предложения, он не может выделить из контекста вспомогательную конструкцию: формы конвербов и связки в любом случае

получают независимые разборы. Иными словами, лемматизатор полностью игнорирует морфосинтаксис. Как и в предыдущих случаях, для поиска собственно аналитических конструкций можно использовать косвенные возможности контекстных запросов.

Отсутствие контекстной информации также, и даже в еще большей степени, сказывается на семантической информации об употреблении тех или иных форм. И в результатах поиска, и в библиотеке в окошке подсветки отображается несколько вариантов английского перевода данного слова. При этом переводы никак не соотнесены с контекстом - пользователю приходится самому выбирать нужный, опираясь на смысл всего предложения. При поиске по переводам возможен точно такой же шум, что и при поиске по грамматическим признакам, лемме или словоформе. Так, поиск по значению 'pit' (яма) даст многочисленные контексты со словоформой *հոր hor*, которая в большем числе случаев, конечно, является родительным или дательным падежом от *հայր hayr* 'отец'. Точно так же лемматизатор не в состоянии определить, в каком значении употреблена та или иная граммема - например, различить реципиентные и локативные употребления дательного падежа или пассивное и реципрокальное употребление медиопассива. В отличие от межлексемной омонимии словоформ (типа обсуждаемых выше *գիր gir* → *գրում grum* ← *գրել grel*) варианты значения лексемы или граммемы трудно различить даже используя контекстный поиск.

Таким образом, при работе с корпусом надо иметь в виду, что в результатах поиска по некоторым запросам может содержаться большое количество ненужных пользователю примеров. Это связано с тем, что лемматизатор анализирует словоформу в отрыве от контекста. Если причина появления "лишних" контекстов на первый взгляд неясна, можно навести на найденную словоформу мышку и посмотреть, какие разборы ей приписаны. После этого в некоторых случаях удается сузить запрос исключением омонимичных разборов или использованием контекстных ограничений, повышающих вероятность искомого явления и/или снижающих объем шума. Однако, как мы видели выше, некоторая вероятность появления шума остается, а некоторая часть нужных примеров при этом может теряться. Пользователь корпуса часто вынужден просматривать найденные контексты, отбрасывая ненужные и сохраняя нужные, правильные результаты поиска.

**Целевая аудитория ВАНК**

Как уже говорилось, аудиторией проекта является в первую очередь сообщество арменистов, работающих с лексикой и грамматикой ашхарабара, а также решающие сравнительные задачи специалисты по западноармянскому языку или грабару. С точки зрения представленности различных форм и временных срезов, ВАНК покрывает значительную часть всего разнообразия языкового материала и ограничен почти только рамками логически невозможных (несовременные устные тексты) или крайне труднодоступных (жанр частной переписки) типов текстов. Как

было сказано, корпус позволяет искать не только определенные словоформы, но и лексемы, и грамматические категории, а также сочетания нескольких слов с определенными признаками в одном контексте (подробнее см. приложение о функциональности). Такого рода функциональность лучше всего приспособлена для изучения лексики и морфологической семантики (значений и правил употребления тех или иных грамматических форм) языка, в определенной степени морфосинтаксиса (например, для изучения глагольных управлений). Синтаксические явления, такие как структура именной группы, стратегии релятивизации, механизмы поддержания референции могут изучаться лишь косвенно, и исследователь-синтаксист должен быть готов применять обходные методы получения релевантных контекстов и/или вручную отсеивать значительное количество поискового "шума".

Благодаря включению переводов и глоссированного формата выдачи 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. Студент-лингвист, таким образом, 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. Студент-нелингвист сможет познакомиться с особенностями значения тех или иных лексем через исследование контекстов их употребления. В принципе, корпус поможет ему составить представление о функционировании тех или иных грамматических форм, но на практике такая постановка вопроса в большей степени свойственна людям с лингвистическим фундаментом.

В целевую аудиторию корпуса, как мы надеемся, могут войти школьные и вузовские преподаватели армянского языка и преподаватели армянского языка как иностранного. Использование корпусов в преподавании - вполне активная, а количественно едва ли не доминирующая отрасль корпусной лингвистики. В последнее время развернута активная пропаганда такого использования Национального корпуса русского языка - осенью прошлого года прошла посвященная этой теме конференция в Москве, по материалам конференции издан сборник статей [Добрушина ред. 2007], имеется также серия публикаций Н. Р. Добрушиной (например, [Добрушина 2005]), начинается разработка образовательного портала корпуса.

Бурное развитие этого направления носит, конечно, вполне прагматический характер (в случае языков, имеющих государственный статус, число изучающих такие языки студентов, по-видимому, всегда больше, чем академических исследователей) и поэтому пропорционально объему преподавания. Для армянского языка этот объем меньше, чем для японского, английского или русского. Но в том объеме, в каком армянский язык преподается, ВАНК мог бы послужить преподавателям важным подспорьем. Он дает возможность работать с живым языковым материалом и отойти от традиционных методов обучения, опирающихся на закрытый и ограниченный объем признанной литературной

классики и жестко нормативного языка. Из корпуса мы узнаем, как на языке говорят, как его в действительности используют.

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. Тем не менее нельзя не признать, что государственное регулирование языковых практик, по крайней мере в сферах, смежных с официальной, является социальной необходимостью и нуждается во внимании со стороны лингвистов. Как можно использовать корпус в языковом планировании? Понимая, что корпус ни в коем случае не является образцом нормы, еще раз повторим, что все-таки именно корпус, представляя действительный языковой узус, может и должен становиться основой для работы над нормой. Языковые реформы, которые оторваны от живых языковых процессов, обречены на фиаско, как видно сегодня на примере бесплодных попыток достичь консенсуса по реформе русской орфографии, а если таковые реформы принимаются, то в конечном итоге они будут отторгнуты языковой стихией В здоровом социуме лингвистический произвол в языковом реформировании невозможен, 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 почти 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. Социолингвисты в основном работают с современным состоянием языка и составляют собственные микрокорпуса, ориентированные на частные задачи. А вот на 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 включение значительного архива армянской периодики, более менее равномерно покрывающего весь период ее существования.

Наконец, часть потенциальной аудитории составляют люди, для которых обращение к корпусу вызвано не профессиональной необходимостью, а личным интересом к языковому материалу. Языковая рефлексия, рассуждения об узусе, значении тех или иных слов характерны для интеллигентного человека вообще. Поэтому при разработке интерфейса ВАНК мы пытались сделать его

функциональность по возможности интуитивной и прозрачной, а разъяснения о том, как искать лингвистическую информацию, максимально неспециальными и свободными от лингвистической терминологии. Пользователь корпуса - нелингвист может искать редкие слова (*եղերդ eġerd* 'цикорий'), слова, в значении которых он сомневается (*բազմապատկիչ bazmapatkič‘* 'множитель'), формы, которые кажутся ему неправильными, но которые он встретил в живой речи или в тексте или запрещаемые нормой формы, которые кажутся ему естественными или допустимыми (*տղու tġu* GEN от 'мальчик'). Привлечение такого рода пользователей требует особых усилий по популяризации корпуса: пользователь - нелингвист, даже если он случайно попадет на сайт корпуса, не сразу поймет, какую пользу он может извлечь из этого инструмента.

**Краткий обзор истории проекта и его перспективы**

Проект ВАНК был запущен в январе 2006 г. по инициативе группы московских исследователей и компании CorpusTechnologies. Летом 2007 г. открыт портал www.eanc.net, на котором размещен первый релиз корпуса. Второй релиз был размещен на том же портале весной 2008 г. Второй релиз отличался от первого объемом поискового корпуса (около 90 млн. словоупотреблений вместо 60 млн. в первом релизе), открытием полнотекстового доступа к более чем ста произведениям армянской классики (Электронная библиотека ВАНК), а также относительно немногочисленными, но важными функциональными расширениями: в разметку добавлены переводы лексем; добавлен специальный "глоссированный" формат отображения текстов, предназначенный для пользователей, не владеющих или недостаточно хорошо владеющих армянским языком; появилась возможность поиска специфической для армянского языка так называемой "внутренней" пунктуации. На момент сдачи статьи в печать (март 2009 г.) готовится к запуску третий релиз, который будет отличаться от второго в первую очередь объемом (около 110 млн. словоупотреблений).

Как кажется, по полноте и репрезентативности литературного языка ВАНК приблизился к некоторому качественному порогу, преодоление которого не только затруднительно, но и не очень осмысленно. Конечно, можно бесконечно продолжать добавлять в корпус те или иные не попавшие в него художественные произведения или публицистику, но качественных улучшений это уже не принесет. Говоря нестрого, для литературного восточноармянского языка корпус ВАНК является вполне репрезентативным. Устный корпус ВАНК не только можно, но и нужно расширять (теоретически до бесконечности), но уже сейчас он входит в число самых крупных устных корпусов в мире и является единственным устным корпусом такого объема для "среднего" языка. Возможное осмысленное расширение проекта - включение принципиально новых текстов на армянском языке в широком смысле этого слова - литературных западноармянских, средне- и древнеармянских текстов. В настоящий момент на базе проекта ВАНК силами нескольких аспирантов в Ереване осуществляется сбор тестовых диалектных корпусов. Ш. Асильбекян работает в Арцваберде, Г. Мкртчян в Шенаване, С.

Давтян в Гусане; для каждого из диалектов планируется собрать около 100 тыс. словоупотреблений.

В некоторых мелких доработках нуждается грамматическая модель, на которую опирается корпус – например, включение в разметку периферийных и пока не анализируемых грамматических категорий (например, усеченные звательные формы). Технически было бы важно оптимизировать некоторые типы запросов: например, запросы с отрицанием, обработка которых на настоящий момент занимает значительное время. Характерный для армянского языка высокий уровень грамматической омонимии наталкивает на мысль о необходимости работы по снятию омонимии или хотя бы о создании подкорпуса со снятой омонимией (ср. подкорпус с вручную снятой омонимией в Национальном корпусе русского языка). Создание синтаксической модели и внесение в корпус синтаксической разметки могло бы резко увеличить число академических и прикладных областей применимости корпуса. Однако добавление автоматического синтаксического парсера требует огромной теоретической разработки и не может быть реализовано в обозримом будущем.

## Источники:


1. Агаян Э. Б. 1976. Արդի հայերենի բացատրական բառարան [Толковый словарь современного армянского языка]. Т. 1-2. Ереван.
2. Галстян Е. Г. (ред.) 1985. Հայ-ռուսերեն բառարան [Армяно-русский словарь]. Ереван.
3. Гргеарян А. К., Арутюнян Н. М. 1987-1989. Աշխարհագրական անունների բառարան [Словарь географических названий]. Ереван.
4. Гюрджинян Д. С. 2005. Անուն խոսքի մասերի թվի կարգը արդի հայերենում. Քերականական բառարան-տեղեկատու [Категория числа имен в современном армянском. Словарь-справочник]. Ереван.
5. Гюрджинян Д. С., Экекян Н. А. 2007. Հայերենում գործածվող տառային հապավումների բառարան [Словарь. Инициальные аббревиатуры в армянском языке]. Ереван.
6. Добрушина Н. Р. 2005. Как использовать Национальный корпус русского языка в образовании? // Национальный корпус русского языка: 2003 – 2005. Результаты и перспективы. М. – 308-330.
7. Добрушина Н. Р. (ред.) 2007. Национальный корпус русского языка и проблемы гуманитарного образования. Теис.
8. Зализняк А. А. 1977. Грамматический словарь русского языка. Словоизменение. М.
9. Хуршудян В. Г., Подлесская В. И. 2006. Армянское *ban* как дискурсивный маркер речевого сбоя // Армянский гуманитарный вестник № 1. Ереван. – 21-42.
10. Хуршудян В. Г. 2006. Средства выражения хезитации в устном армянском дискурсе в типологической перспективе. Диссертация ... кандидата фил. наук. М.: РГГУ.
11. www.armenianhouse.org – армянская электронная библиотека
12. www.eanc.net - Восточноармянский национальный корпус
13. http://forum.am-kayq.com – армянский форум
14. www.hayeren.hayastan.com – армянский образовательный портал
15. http://www.linguistlist.org/sp/texts.html – Страница ресурса «Linguistlist», посвященная электронным корпусным ресурсам.
16. www.ruscorpora.ru - Национальный корпус русского языка
17. www.sd-editions.com/LALT/home.html – Leiden Armenian Lexical Textbase
18. http://titus.uni-frankfurt.de/indexe.htm – Thesaurus Indogermanischer Text- und Sprachmaterialien

*Приложение*

Краткая характеристика функциональности ВАНК

ВАНК ориентирован в первую очередь на лексические и грамматические запросы. Синтаксические запросы можно делать лишь опосредовано, так как корпус не имеет синтаксической разметки. По поисковой функциональности ВАНК очень близок Национальному корпусу русского языка (www.ruscorpora.ru), который в значительной степени служил его прототипом. В рамках настоящей приложения поисковая функциональность может быть описана лишь в самых общих чертах, для более подробного ознакомления мы приглашаем читателя зайти на сайт www.eanc.net. Итак, возможны следующие типы поиска.

*1. Поиск словоформы или лексемы.* ВАНК позволяет искать как вхождения конкретной словоформы (например, *մարդու mardu*), так и вхождения всех словоформ определенной лексемы (например, словоформ *մարդ mard*, *մարդու mardu*, *մարդիկ mardik* и т.д. от лексемы *մարդ mard*).

*2. Поиск по переводу*. 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 быть определен как логическая конъюнкция или дизъюнкция нескольких категорий или совмещать конъюнкцию и дизъюнкцию в одной логической формуле.

*4. Дополнительные признаки*. 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.

*5. Контекстные запросы*. ВАНК позволяет искать сочетания конкретных словоформ, лексем или словоформ, определенных грамматическими признаками. Иными словами, все описанные выше "точечные" запросы на поиск одного слова можно сочетать в контекстные запросы, которые ищут сочетания этих точечных запросов в одном контексте. Расстояние между вхождениями можно изменять, меняя интервал допустимых расстояний. ВАНК позволяет искать вхождения не в одном, а в соседних (и далее) предложениях.

*Выбор подкорпуса*. Любой запрос, который может быть применен к ВАНК, может быть также применен и к определенному пользователем подкорпусу ВАНК. Окно подкорпуса состоит из следующих зон: авторы и произведения, период, жанр

(с достаточно подробной классификацией жанров и типов текстов), проза/поэзия, оригинальные / переводные тексты, детская / общая литература. Эти признаки можно использовать одновременно, например, ограничивая поиск стихотворными текстами первой половины двадцатого века и т.п.

*Отображение результатов***.** ВАНК позволяет осуществлять сортировку контекстов по целому ряду параметров: лемма (исходная форма), словоформа, словоформа слева от вхождения, автор, название произведения, автор, год создания, 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.

*4. Глоссированный*: этот формат предназначен в первую очередь для лингвистов-типологов и людей, начинающих изучать армянский язык. Данное 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 в виде столбца, расположенного непосредственно под словоформой, 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 приводятся грамматические (словоизменительные) признаки словоформы. Если лексеме приписаны переводы, они даются в третьей строчке. Разные разборы одной словоформы визуально отделены друг от друга.

Каждый контекст представлен в окне выдачи одним предложением; слова-вхождения при этом выделены цветом. При контексте приводятся базовые библиографические характеристики: 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).

Результаты можно отображать как в армянском алфавите, так и в латинской транслитерации. Используемая в ВАНК транслитерация в основном следует международной армяноведческой традиции Хюбшманна-Мейе, адаптированной

под Unicode. Транслитерация используется в том числе при отображении имени автора и названия произведения.

При выдаче результатов запроса в верхней части экрана отображается общая информация об отклике и исходном запросе, в том числе:

1. Число вхождений (в случае больших контекстных запросов – примерная оценка их общего числа в корпусе).
2. Число документов (в случае больших контекстных запросов – примерная оценка числа документов, в которых они могут встретиться).
3. Выбранные критерии сортировки и размер подкорпуса, по которому осуществлялся поиск.

Ниже приводятся некоторые примеры микрозадач, которые могут решаться при помощи корпуса. Соответствующая каждой задаче последовательность действий (что нужно делать, чтобы ее решить) подробно описана на сайте в разделе "Как искать (примеры запросов)".

1. Найти вхождения словоформы *տղու tġu*
2. Найти все вхождения всех словоформ существительного *էրէրդ eġerd*
3. Найти формы императива множественного числа от медиопассивных глаголов
4. Найти формы императива глагола *անձրևել anjrevel*
5. Найти все словоформы лексем, заканчивающихся на *-թյուն -t‘yun* и начинающихся на *ան- an-*
6. Сравнить использование глагола *պրծացնել prcac‘nel* в 19-м веке с его использованием в 20-м веке и сегодня.
7. Сравнить использование глагола *պրծացնել prcac‘nel* в поэзии 19-го века с его использованием в поэзии 20-го века.

*Статистика использования словоформ.* В готовящемся на настоящий момент к запуску 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. Ниже в качестве примера приведена таблица со статистикой употребления словоформы *աստծո astco* (родительный падеж единственного числа от нерегулярно склоняющегося существительного *աստված astvac* 'бог') по данным корпуса. 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.

**Словоформа: *ши ин бп.* Число вхождений: 9,195 Ранг: 386.9 WPM: 548.**

| | **Художественные** | **Нехудожественные** | **Пресса** | **Устные** | **Всего по декаде** |
|---|---|---|---|---|---|
| (1800-1859) | 18 | n/a | 0 | n/a | 11 |
| (1860-1869) | 95 | n/a | 0 | n/a | 61 |
| (1870-1879) | 108 | 40 | 0 | n/a | 74 |
| (1880-1889) | 90 | 124 | 0 | n/a | 70 |
| (1890-1899) | 72 | n/a | n/a | n/a | 72 |
| (1900-1909) | 56 | 39 | 0 | n/a | 28 |
| (1910-1919) | 179 | n/a | 0 | n/a | 121 |
| (1920-1929) | 14 | 0 | 3 | n/a | 10 |
| (1930-1939) | 75 | 8 | 4 | n/a | 45 |
| (1940-1949) | 77 | 60 | 0 | n/a | 47 |
| (1950-1959) | 64 | 39 | 8 | 187 | 47 |
| (1960-1969) | 258 | 45 | 14 | 318 | 171 |
| (1970-1979) | 126 | 45 | 6 | 55 | 79 |
| (1980-1989) | 224 | 24 | 11 | 182 | 91 |
| (1990-1999) | 182 | 57 | 91 | 32 | 112 |
| (2000-2009) | 169 | 127 | 100 | 61 | 59 |
| недатированные | 171 | 17 | 388 | 456 | 161 |
| **Всего по жанру** | **151** | **55** | **38** | **68** | |