

УДК 811.51

**НАЦИОНАЛЬНЫЙ КОРПУС КАЗАХСКОГО ЯЗЫКА
КАК ИННОВАЦИОННО-ИНФОРМАЦИОННАЯ БАЗА
ГОСУДАРСТВЕННОГО ЯЗЫКА
(ОБ ОСОБЕННОСТЯХ И ЗАДАЧАХ ПЕРВОГО
ЭТАПА РАЗРАБОТКИ И СОЗДАНИЯ)**

Анар Фазылжанова

к.ф.н.

зам. дир. Института языкознания имени Ахмета Байтурсынулы

Аннотация

Национальный корпус казахского языка в данное время еще не решен полностью ни в теоретическом, ни в практическом планах. Однако есть в наличии отдельные исследовательские проекты по актуальным проблемам создания корпуса казахского языка, выполненные Институтом языкознания им. Ахмета Байтурсынулы, и научные публикации отдельных ученых лингвистов, таких как профессор Сулейменова Э.Д., посвященные проблемам его создания в последнее десятилетие. При разработке конкретных параметров Национального корпуса языка считаем необходимым на самых ранних стадиях его теоретического обоснования иметь четкое понимание о приоритетах создания корпуса вообще, о глубине и ширине Национального корпуса казахского языка, его отличиях и своеобразии от других языковых корпусов.

В данной статье рассматриваются важные приоритеты создания Национального корпуса казахского языка, а также задачи первого этапа, которые позволят вывести самые главные особенности корпуса казахского языка.

Ключевые слова: национальный корпус, казахский язык, информационная база, инновация и т.д.

**КАЗАК ТИЛИНИН УЛУТТУК КОРПУСУ – МАМЛЕКЕТТИК ТИЛДИН
МААЛЫМАТ БАЗАСЫ КАТАРЫ**

Қысқаша мазмуну

Бұл мақала қазақ тилинин корпусун түзүүгө арналган.

Ачкыч сөздер: улуттук корпус, қазақ тили, маалымат базасы, инновация.

THE NATIONAL CORPUS OF THE KAZAKH LANGUAGE AS THE INNOVATIVE AND INFORMATION BASE OF THE STATE LANGUAGE

Abstract

The National Corpus of the Kazakh language has not been solved completely either in theoretical or in practical aspects at present yet. However, there are available separate research projects on actual problems of the creation of the Corpus of the Kazakh language executed by the Institute of Linguistics named after Akhmet Baitursynuly and scientific publications of certain scientific linguists, such as Professor Suleymenova E.D., devoted to the problems of its creation in the last decade. When developing concrete parameters of the National Corpus of the language we consider it necessary at the earliest stages of its theoretical justification to have a clear understanding about the priorities of creation of the Corpus in general, about the depth and width of the National Corpus of the Kazakh language, its differences and originality from other language corpora.

In this article, important priorities of creation of the National Corpus of the Kazakh language, and the problems of the first stage which will allow to reveal the most important features of the corpus of the Kazakh language are considered.

Key words: National Corpus, the Kazakh language, information base, innovation.

На сегодня тема, поднятая в данной статье, может показаться неуместной, так как сам вопрос создания и разработки Национального корпуса казахского языка еще не решен полностью ни в теоретическом, ни в практическом планах. Хотя не следует отрицать наличие отдельных исследовательских проектов по отдельным проблемам создания корпуса казахского языка (Научно-исследовательские проекты отдела “Прикладная лингвистика” Института языкознания “Компьютерная база казахского языка: открытая система лексического фонда” (2006-2008 гг.), “Корпус текстов казахского языка как культурная ценность и национальный компьютерный фонд словарей “Тіл – казына” (2009-2011гг.), “Аннотированный корпус казахского литературного языка” (2013-2014 гг.), “Принципы и разработка метатекстовой разметки в национальном корпусе казахского языка” (2015 г.)), выполненных Институтом языкознания имени Ахмета Байтурсынова и научных изысканий, ученых лингвистов, таких как профессор А.К. Жубанов (Жубанов, 2012) профессор Э.Д. Сулей-

менова, посвященных теоретическим и прикладным вопросам его создания (Сулейменова, 2011). Однако же при разработке конкретных параметров Национального корпуса языка считаем необходимым на самих ранних стадиях его теоретического обоснования иметь четкое понимание о приоритетах его создания вообще, о глубине и ширине Национального корпуса казахского языка, отличиях и своеобразии от других языковых корпусов. Для того чтобы был понятен полный смысл последнего высказывания, хотим остановиться на вышеупомянутых понятиях и основных категориях НККЯ, таких как приоритеты, глубина и ширина, отличие и своеобразие.

О ПРИОРИТЕТАХ. Лингвисты могут возразить даже самой постановке этой проблемы, так как известно, что во всем мире единственным неписанным приоритетом создания языковых корпусов всегда было только одно, это – улучшение качества лингвистических исследований, расширение их эмпирической базы посредством создания огромных массивов глубоко аннотированных текстов. Однако

же опыты и языковые ситуации создания корпусов развитых стран намного отличаются от казахстанского. На сегодня известно, что самые объемные языковые корпусы были созданы в странах с развитой инфраструктурой и информационной технологией, и еще, интересен тот факт, что именно в странах с моноязыковой ситуацией и развитыми государственными языками (Япония, Англия, Россия, Финляндия и т.д.). По сравнению с этими странами Казахстан отличается полиязыковой ситуацией, а также немногочисленным опытом развития собственного государственного языка. К тому же, не умаляя достоинств интенсивного экономического развития, следует признать, что в стране мало развиты информационные технологии (инфраструктура национального лингвистического корпуса), связанные с лингвистическими инновациями, а именно с программированием казахского языка (нет специальности на филологических факультетах именно по корпусной лингвистике, IT-специалистов с профессиональным знанием казахского языка, создающих языковые компьютерные программы). В связи с этим при создании корпуса казахского языка лингвистам необходимо учитывать не только интересы чисто научного характера, т.е. потребности в корпусе ученых-лингвистов, а также потребности пользователей корпуса – простых граждан Республики Казахстан. Поэтому для корпуса казахского языка вопрос о правильной постановке приоритетов немаловажен. Анализируя современную ситуацию, сложившуюся вокруг казахского языка, проблемы, связанные с расширением сфер его употребления и улучшением качества функционирования, считаем, что первым приоритетом создания НККЯ должно быть укрепление роли государственного казахского языка как идентификатора общегосударственной национальной идентичности, т.е. укрепление базы государственного языка так, чтобы он смог

стать мощным консолидирующим фактором для казахстанского поликультурного и полиязыкового сообщества. Как это осуществить? Для этого, во-первых, необходимо на самых первых временных этапах (нами этапы создания НККЯ корпуса разделены на первый, второй и третий с продолжительностью в 3 года и задачей охвата текстов на каждом этапе по 100 миллионов словоупотреблений, в итоге, за 9 лет необходимо создать НККЯ объемом в 300 миллионов словоупотреблений) создания корпуса отборочно относиться к текстовым материалам. Сам по себе корпус должен быть адекватным отражением языкового существования во всем его многообразии независимо от стилей и жанров, подстилей и социальных оттенков речевого воспроизводства авторов текста и т.д. Однако, взяв это под общую цель, на начальном этапе формирования базы текстов корпуса, который будет сразу же после разработки представлен потребителю в интернете, следует отдать большее предпочтение культурно значимым текстам, с содержательным контентом, более того, патриотического характера, будь то в подкорпусе художественных текстов, в газетном подкорпусе и др. подкорпусах. Под текстами культурной значимости мы вслед за Д.В.Сичиновой, ученым-лингвистом, одним из активных создателей национального корпуса русского языка, понимаем “языковой пример, освященный авторитетом сочинителя” (Сичинова, 2002: 26). По этому поводу ученый писал: “Сколько бы не декларировалась независимость лингвистики от эстетико-культурного подхода, письменный язык – это прежде всего язык культуры. От лингвистической работы часто требуется не языковой пример, а (сколь угодно субъективно понимаемый) хороший языковой пример, “освященный” авторитетом сочинителя; характерно, что такое требование возникает во всех лингвистических традициях. Поэтому собрание культурно значимых

текстов на данном языке также представляет собой обладающий собственной ценностью источник для аннотированного корпуса”, с чем мы полностью согласны (Сичинава, 2002: 26). Опыт создания культурно-репрезентативных корпусов имеется в европейских странах. Так, например, известно, что созданный Австрийский национальный корпус в Вене включает тексты на немецком языке, оказавшие определенное влияние на национальную культуру (Сичинава, 2002: 26).

Во-вторых, по мере формирования финансовой, а также научно-материальной и уже самой корпусной (текстовой) базы приступить к осуществлению и созданию корпусных ресурсов параллельных подкорпусов казахско-русского, казахско-английского, а также казахско-турецкого, казахско-узбекского, казахско-киргизского и других языков диаспор, проживающих на территории Республики, что усилит возможности и научно-лингвистические основы укрепления статуса казахского языка как языка межэтнического общения, тем самым способствуя укреплению общегосударственной идентичности.

В-третьих, считаем, что необходимо увеличить общепринятый в корпусной лингвистике объем энциклопедической информации не только по количеству параметров метаразметки, но и особым видам экстралингвистического аннотирования корпусных текстов. Для ясности, разясним, что на сегодня в корпусной лингвистике под энциклопедической понимается информация, не касающаяся лингвистических характеристик слова, т.е. языковых знаний. В основном в современных Национальных корпусах такой вид информации передается через особый вид аннотирования как метаразметка или метаданные, включающие в себя внешнюю, интеллектуальную разметку (библиографические характеристики, типологические характеристики, тематические характеристики, социологические

характеристики), “формальную” структурную разметку (текст, раздел, глава, абзац, предложение), а также технико-технологическую разметку (кодировку, даты, обработки, исполнителей, источник электронной версии), набор которых во многом определяет возможности, предоставляемые корпусами исследователям (см. об этом: Лекция 4). Но речь идет не о метаразметке. Хотя она несет в себе энциклопедическую информацию. Мы хотим обратить внимание на особый вид аннотирования корпусных текстов как культурно-семантическая разметка (термин употреблен впервые нами), которую можно отнести к аннотации корпуса по энциклопедическим видам информации также как и метаразметку. Однако она сильно отличается от метаразметки и по методу подачи в корпусе, и по своему содержанию. Результаты изученных нами материалов по созданным корпусам показывают, что в них не предпринималась попытка подачи культурно-семантической информации, если даже и были попытки прямой (а не косвенной) передачи данного типа семантики, то нам еще не известны результаты таковых, и этот факт все более убеждает нас в том, что данный тип аннотирования все же еще не разработан до завершенной системной формы. Так вот, несмотря на такое положение дел, т.е. отсутствие какого-либо опыта, да еще в самом начале пути создания корпуса, когда еще не разработаны полностью параметры его глубины и ширины, говорить о культурно-семантической разметке может показаться дерзким, возможно даже авантюрным решением. Все же мы намерены предпринять такую попытку не столько из чисто профессиональных побуждений, как “создать улучшенный корпус”, “отличиться от других”, “пойти дальше”, а сколько из намерений послужить требованиям времени и ответить запросам общества “здесь и сейчас”. Разясним: на сегодня в казахоязычном пространстве и сообществе наряду с такими

положительными трендами как рост числа говорящих, словотворческий, терминотворческий бум, улучшение качества использования языка в сферах и каналах ранне по внешним факторам искусственно закрытых и запрещенных, интенсивное увеличение языковых инновации, наблюдается мощный отрицательный тренд потери “культурной памяти” народом, “стирания культурных кодов” в сознаний, которые заключены в глубинных семантиках казахских слов. Поэтому раскрытие этих культурных кодов, сокрытых в этномаркированных (лингвокультурологических, неэквивалентных, лингвокогнитологических и др.) языковых единицах имеет не только чисто научную значимость, а также высокое практическое значение для казахской этноязыковой общности в целях цементирования его идентичности, а также сохранения духовной безопасности и укрепления иммунитета национальной культуры и духа против вестернизации, девальвации духовных ценностей. И именно корпус как инновационно-информационная открытая система, доступная в онлайн-режиме, предоставляющая пользователю массив казахских аннотированных текстов, был бы мощным и эффективным проводником таких – лингвокультурологических знаний, сокрытых в культурной семантике языковых единиц. Надо признать, что идея подачи культурно-семантической разметки, несмотря на свою привлекательность в плане обогащения знаний пользователя НККЯ о глубинных смыслах казахских слов, пока еще остается не разработанной в техническом аспекте. В лингвистическом аспекте подача лингвокультурологической дефиниции, разработанная в технологиях Большого однотомного толкового словаря “Қазақ сөздігі” и 15-томного толкового словаря “Қазақ әдеби тілінің сөздігі”, выпущенных под идейным и научным руководством профессора Н.Уалиева, в общем, приведена в более менее системный вид и представ-

лена как опыт “Лингвокультурологического словаря” в данном исследовании, который будет использован в НККЯ. На этот опыт мы полагаем опираться в подаче культурно-семантической разметки в Корпусе. Однако, следует принять во внимание тот факт, что этот опыт – первичный, и, наверное, примитивный, а также создание корпуса, скорее не только научный, а более того – творческий проект. Ввиду этого хотим предупредить, что выработанные нами технологии раскрытия культурной семантики и методы его дефиниции могут по мере разработки корпуса совершенствоваться, улучшаться и меняться.

Таким образом, на данном этапе исследования нами предлагается минимодель – примитивная, пилотная версия корпуса казахского языка, в которой предпринята попытка подачи культурно-семантической разметки. Она будет находиться в системе поиска в разделе “семантика”, которая в свою очередь будет состоять из “общей семантики”, где дается лексическая и грамматическая, т.е. лингвистическая семантика в целом, и “культурной семантики”, где раскрывается культурный код (если таковой присутствует в слове), сокрытый в данной языковой единице.

О ГЛУБИНЕ И ШИРИНЕ НККЯ. Под глубиной Национального корпуса подразумевается охват текстов разных хронологических отрезков. И чем богаче хронологические периоды, тем более глубинным будет корпус, который позволит исследователю тщательно проследить историю языковых процессов, единиц и явлений. Сюда же можно отнести и охват текстов в разных графических системах. Поэтому на начальном этапе создания НККЯ предполагается охватить тексты в трех графиках: на тоте жазу (казахская арабица начала XX века), латинице (1929-1941 гг.) и кириллице. Таким образом, глубина НККЯ на начальном этапе будет охватывать 2 века: XX и XXI. А по

усилению возможностей, укреплению технической оснащенности и увеличению объема в последующих этапах необходимо создать подкорпус исторических текстов, начиная с древних и средневековых письменных памятников. Также под глубиной языковых корпусов иногда понимают оснащенность корпуса разметками. “Под разметкой (англ. annotation) понимается содержащаяся в корпусе лингвистическая или иная информация, приписанная тем или иным отрезкам текста (так, информация о времени создания и жанре приписывается целому тексту, информация о синтаксической структуре – предложению, информация о лексеме и грамматических характеристиках – слову), а равно и процесс, добавляющий к исходному тексту эту информацию” (Сичинава, 2002: 27).

Нам представляется возможным на первом этапе оснастить НККЯ такими видами разметки как: морфологическая, семантическая (лексико-грамматическая семантика), культурно-семантическая (лингвокультурологическая) и метаразметка, на последующих этапах: просодическая, стилистическая и синтаксическая.

Ширина языкового корпуса – понятие, охватывающее содержание корпуса, такие его параметры как представительность и сбалансированность. Сам по себе объемный массив оцифрованных текстов, даже аннотированных, не представляет высокой значимости для исследователя, если он не сбалансирован, не представлен. “Национальный корпус (Национальный корпус русского языка (НКРЯ) – А.Ф.) характеризуется представительностью, или сбалансированным составом текстов. Это означает, что создатели стремятся включить в Корпус все типы письменных текстов, представленные в языке (художественные разных жанров, публицистические, учебные, научные, деловые, эпистолярные и т.п.), и что все эти тексты входят в корпус про-

порционально их доле в языке соответствующего периода. Так, среди текстов XIX века преобладают художественные, а среди современных – публицистические и, во вторую очередь, научные, при этом в Основной корпус включены и эпистолярные тексты, и тексты, представляющие различные жанры современного интернет-общения. Конечно, идеальное подобие реальному языку создать невозможно, но все же при принятии решений о количественном соотношении жанров в Корпусе для каждого временного периода решающим фактором становится стремление максимально приблизиться к естественному существованию языка” (Образовательный портал Национального корпуса русского языка: сайт). Соотношение разнотипных текстов, входящих в Основной корпус (русского языка – А.Ф.), описано на странице “статистика” НКРЯ (Национальный корпус русского языка: сайт).

ОБ ОТЛИЧИЯХ И СВОЕОБРАЗИИ. Таким образом, вышеназванные приоритеты создания НККЯ, а также задачи первого этапа, рассмотренные здесь, позволяют вывести такие особенности корпуса казахского языка: во-первых, содержание текстовых ресурсов НККЯ на первом этапе будут иметь культурную значимость, а сам корпус отличаться культурно-репрезентативным характером; во-вторых, предполагается создать первый Национальный языковой корпус с культурно-семантической аннотацией на базе казахского языка, в-третьих, на втором и третьем этапе разработки НККЯ предполагается усилить его значимость и ценность созданием и расширением ресурсов параллельных подкорпусов казахского языка с другими языками диаспор, проживающими в Казахстане.

ЛИТЕРАТУРА:

1. ЖҰБАНОВ, Аскар (2012). Қазақ тіл білімі: колданбалы лингвистика. Алматы: “КИЕ”. – 696 б.

2. Лекция 4. Экстралингвистическая разметка. Метаданные. – See more at: <http://keepslide.com/humor/17562#sthash.DaJ80kCZ.dpuf>
3. Национальный корпус русского языка. <http://www.ruscorpora.ru/corpora-stat.html>.
4. Образовательный портал Национального корпуса русского языка. http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=300&Itemid=127
5. СИЧИНАВА, Д.В. (2002). К задаче создания корпусов русского языка”, НТИ. Серия 2 Информационные процессы и системы. №11. 25-31 сс.
6. СУЛЕЙМЕНОВА Элеонора (2011). “Қазақ тілі үшін ұлттық корпус керек пе?”, ҚазҰУ хабаршысы. Филология сериясы, № 1 (131). 76-79 с.;