Мадиева Г.Б., Уматова Ж.М.

Об Алматинском корпусе казахского языка

Современные информационные технологии и технические средства открывают новые возможности для лингвистического исследования на базе языковых корпусов. В статье представлено описание пилотной версии Алматинского корпуса казахского языка (АККЯ), как составляющей Национального корпуса казахского языка (НККЯ), представляющего справочно-информационную систему на основе обширного фонда размеченных текстов литературного казахского языка, созданного в рамках Государственной программы функционирования и развития языков (2011-2020).

Описана история создания, роль, наполненность, размеченность и целесообразность использования АККЯ в лингвистических исследованиях, в практике преподавания казахского языка, как родного и иностранного, в качестве источника для составления лексикографических изданий.

Ключевые слова: языковой корпус, национальный корпус, информационно-справочная система, подкорпус, разметка текстов, прикладное значение корпуса.

Madiyeva G.B., Umatova Zh.M.

About Almaty kazakh language corpus

Modern information technologies and technical means open new opportunities for linguistic research on the basis of language corpora. The description of the pilot version of Almaty Corpus of the Kazakh Language (ACKL), as a component of the National Corpus of the Kazakh Language (NCKL) representing the reference system on the basis of extensive fund of the marked texts of the literary Kazakh language created within the State program of functioning and development of languages (2011-2020) is presented in the article.

The history of creation, role, fullness, markedness and expediency of using ACKL in linguistic researches, in practice of teaching the Kazakh language, as a native and foreign one, as a source for drawing up lexicographic editions is described.

Key words: language corpus, national corpus, reference system, subcorpus, marking of texts, applied value of the corpus.

Мадиева Г.Б., Уматова Ж.М.

Алматы қазақ тілі корпусы туралы

Қазіргі ақпараттық технология және техникалық құралдар тіл білімінде тілдік корпустар негізінде зерттеулер жасауға үлкен мүмкіндік тудырып отыр. Мақалада Қазақ тілінің ұлттық корпусын құраушы қазақ тілінің Алматы корпусының Мемлекеттік тілді дамыту мен жоспарлау бағдарламасы (2011-2020) аясында жасалған ақпараттық-анықтамалық жүйеде қазақ тілінің әдеби тілінің ауқымды фондық мәтіндері берілген. Алматы корпусының жасалу тарихы, рөлі, толықтырылуы, Алматы корпусының тілдік зерттеулер жасаудағы көмегі, қазақ тілін үйрету тәжірибесіндегі маңызы, ана тілі және шетел тіліндегі лексикографиялық басылымдар ретіндегі орны айтылады.

Түйін сөздер: тілдік корпус, ұлттық корпус, ақпараттық-анықтамалық жүйе, корпус іші, мәтіндерге анықтама жасау, корпустың қолданбалы маңызы.

УДК 811.111

Мадиева Г.Б.[1], Уматова Ж.М.[2]

[1]д. ф. н. профессор, [2]к. ф. н. доцент Казахского национального университета имени аль-Фараби, г. Алматы, Казахстан,
e-mail: gmadi@mail.ru, zh.umatova@mail.ru

ОБ АЛМАТИНСКОМ КОРПУСЕ КАЗАХСКОГО ЯЗЫКА

Специалисты различных сфер деятельности: политологи, культурологи, экономисты и, в первую очередь, лингвисты, неоднократно отмечают, что за последние годы казахский язык все больше расширяет свои границы. Так, Ельдесов Д. пишет, что «С приданием казахскому языку статуса государственного его роль в республике значительно усилилась, обеспечивая законодательным правом использоваться во всех функциях и сферах общения. Статусному положению языка, политическим и законодательным мерам должно быть соответственное сугубо лингвистическое наполнение, и в этом плане появилась проблема корпусного планирования – попытки стандартизировать, упорядочить и систематизировать язык. В Республике Казахстан в рамках Государственной программы функционирования и развития языков (2011-2020), необходимо создать Национальный корпус казахского языка. Формирование национального корпуса языка стало актуальной проблемой во многих республиках после распада Советского Союза, поскольку государственному статусу языка могут соответствовать лишь кодифицированные (нормированные) языки» [1].

Ведущий лингвист не только казахстанского, но и мирового сообщества Сулейменова Э.Д. отмечает: «Создание национальных корпусов базовых государственных языков ведущих стран мира возведено в ранг важных историко-культурных и политических мероприятий современности. Большинство крупных языков мира уже имеет свои национальные корпуса (различающиеся по полноте и уровню научной обработки текстов). Общепризнанным образцом является, в частности, Британский национальный корпус.

Создание корпуса позволит изучать историю казахского языка, осуществить статистический мониторинг функционирования лексических, грамматических и стилистических языковых средств, работать по лексикографической поддержке современного казахского языка, его стандартизации, создавать словари, учебники, справочные пособия. Национальный корпус казахского языка способен служить современным источником его кодификации и стандартизации, поскольку в корпусе оказывается зафиксированным письменный и звучащий язык в его максимально репрезентативном виде. Формирование На-

ционального корпуса казахского языка одна из важнейших, задач суверенного Казахстана» [2].

В Википедии лингвистический корпус определяется как «совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой» [3]. Из чего можно заключить, «что национальный корпус казахского языка –

– это информационно-справочная система на базе электронного собрания письменных и звучащих текстов, сбалансированная и представительная по объему (сотни миллионов словоупотреблений), оснащенная всеми возможными видами полной и удобной разметки» (Сулейменова Э.Д.) [2].

Учитывая назревшую острую необходимость, в рамках идеи «Мәңгі ел – Мәңгі қазақ тілі» в Казахском национальном университете имени аль-Фараби в мае 2012 г. при поддержке ректора Г.М. Мутанова началась работа над проектом Корпуса. Силами кафедры общего языкознания и европейских языков факультета филологии и мировых языков под руководством заведующей кафедрой Г.Б. Мадиевой при участии сотрудников факультета филологии Национального исследовательского университета Высшая школа экономики (Москва) Архангельского Т.А., Бонч-Осмоловской А.А., Даниэля М.А., Ляшевской О.Н., Толдовой С.Ю. в мае 2014 года была выпущена пилотная версия Алматинского корпуса казахского языка, представляющего собой интеллектуальную информационно-справочную систему на основе обширного фонда размеченных текстов в удобной для использования современной виртуальной форме [4].

Для корпуса была адаптирована поисковая система Восточноармянского национального корпуса (EANC).

Это первая версия корпуса Национального корпуса казахского языка – НККЯ как справочно-информационной системы на основе обширного фонда размеченных текстов литературного казахского языка, государственного языка Республики Казахстан. В настоящий момент размер корпуса составляет около 1 миллиона словоупотреблений. Тексты корпуса были размечены с помощью автоматического морфологического анализатора, 75 % словоформ корпуса имеют грамматический разбор. Омонимия в корпусе не снималась, т.е. каждой словоформе приписаны все возможные варианты разбора без учета контекста.

Алматинский корпус казахского языка представлен письменными текстами современного казахского языка, которые сбалансированно распределены по нескольким подкорпусам:

– художественный;
– научный;
– художественно-публицистический.

Ключевой особенностью любого корпуса является наличие не только метатекстовой информации, но и дополнительной, так называемой разметки, позволяющей использовать его данные для научных исследований. В Алматинском корпусе в настоящее время используется три вида разметки:

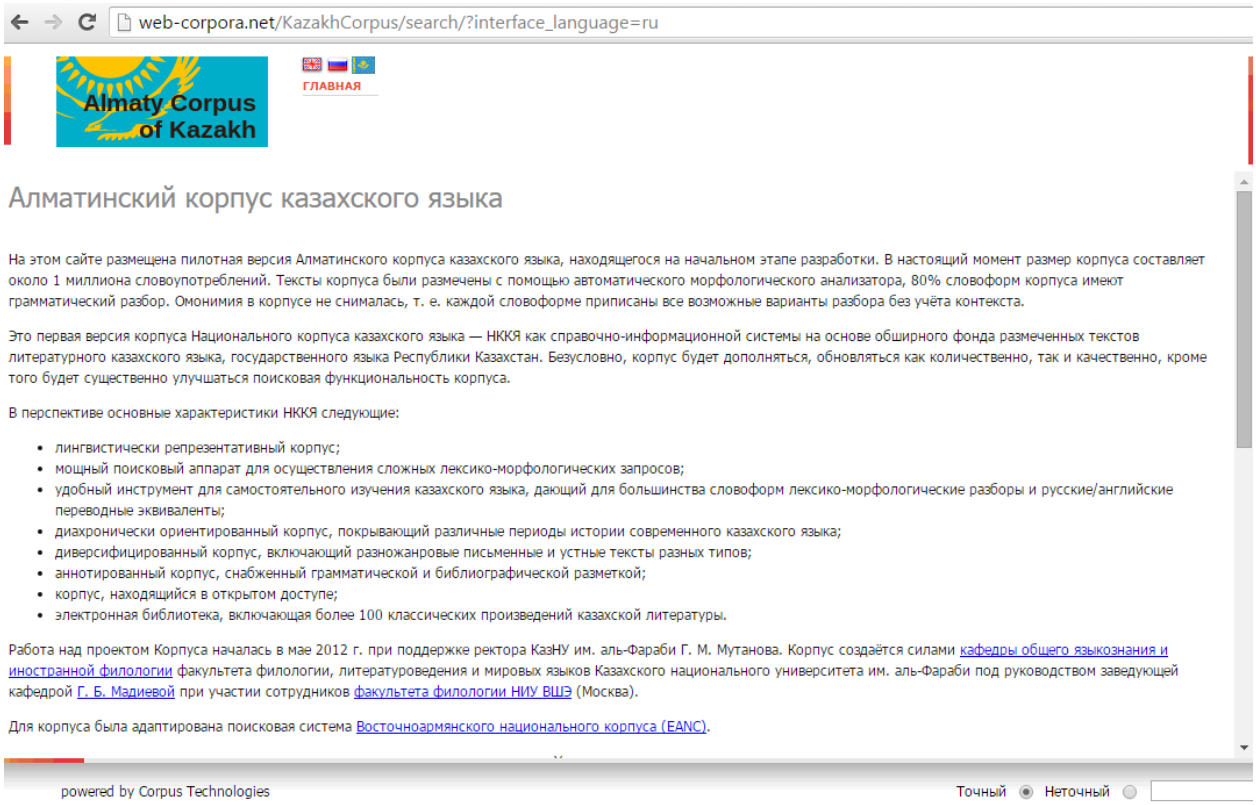
– морфологическая, т.е. частеречная разметка, которая включает не только признак части речи, но и признаки грамматических категорий, свойственных данной части речи. Схема морфологической разметки предполагает наличие, во-первых, набора тэгов каждого казахского слова, попавшего в словарь во-вторых, описания того, что каждый из них означает и, в-третьих, правил присвоения тэгов единицам текста;

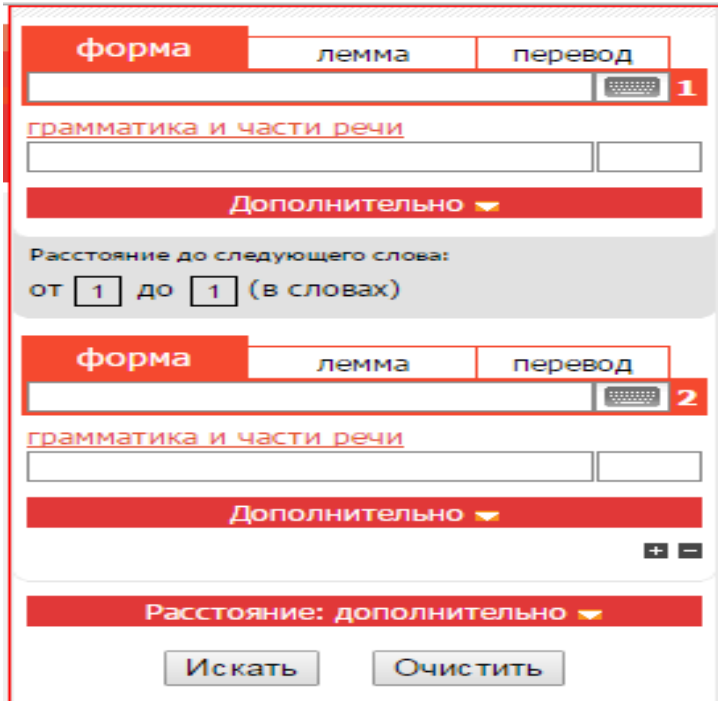
– синтаксическая, как результат синтаксического анализа или парсинга (от англ. parsing), т.е. это грамматика структур непосредственно составляющих;

– семантическая, при помощи специального кода, состоящего из букв и цифр или только цифр, в котором первая буква или цифра обозначает общую семантическую категорию, в которую входит данное слово, а последующие символы – более узкие подкатегории, специализирующие его значение. В схемах семантической разметки предусмотрены те случаи, когда в качестве единицы смысла выступает не отдельное слово, а словосочетание.

Была разработана поисковая система корпуса казахского языка, в которой поиск производится не только по конкретному слову, но и по грамматическим признакам.

К настоящему времени текстовая база электронных произведений представлена казахскими авторами, прежде всего, классиков казахской литературы: Абай, М. Ауэзов, А. Нуршаихов, И. Есенберлин, А. Нурпеисов, Г. Мусрепов, С. Торайгыров, С. Сейфуллин и др.; произведениями классиков мировой литературы, переведенные на казахский язык: Ч. Айтматов, публицистика: газеты Егемен Казакстан, Айкын, Халык сөзі, Ана тілі, Айқын и др.; научные тексты: докторские и кандидатские диссертации, монографии, статьи.


web-corpora.net/KazakhCorpus/search/?interface_language=ru
Almaty Corpus of Kazakh
ГЛАВНАЯ
Алматинский корпус казахского языка
На этом сайте размещена пилотная версия Алматинского корпуса казахского языка, находящегося на начальном этапе разработки. В настоящий момент размер корпуса составляет около 1 миллиона словоупотреблений. Тексты корпуса были размечены с помощью автоматического морфологического анализатора, 80% словоформ корпуса имеют грамматический разбор. Омонимия в корпусе не снималась, т. е. каждой словоформе приписаны все возможные варианты разбора без учёта контекста.
Это первая версия корпуса Национального корпуса казахского языка — НККЯ как справочно-информационной системы на основе обширного фонда размеченных текстов литературного казахского языка, государственного языка Республики Казахстан. Безусловно, корпус будет дополняться, обновляться как количественно, так и качественно, кроме того будет существенно улучшаться поисковая функциональность корпуса.
В перспективе основные характеристики НККЯ следующие:
лингвистически репрезентативный корпус;
мощный поисковый аппарат для осуществления сложных лексико-морфологических запросов;
удобный инструмент для самостоятельного изучения казахского языка, дающий для большинства словоформ лексико-морфологические разборы и русские/английские переводные эквиваленты;
диахронически ориентированный корпус, покрывающий различные периоды истории современного казахского языка;
диверсифицированный корпус, включающий разножанровые письменные и устные тексты разных типов;
аннотированный корпус, снабженный грамматической и библиографической разметкой;
корпус, находящийся в открытом доступе;
электронная библиотека, включающая более 100 классических произведений казахской литературы.
Работа над проектом Корпуса началась в мае 2012 г. при поддержке ректора КазНУ им. аль-Фараби Г. М. Мутанова. Корпус создаётся силами кафедры общего языкознания и иностранной филологии факультета филологии, литературоведения и мировых языков Казахского национального университета им. аль-Фараби под руководством заведующей кафедрой Г. Б. Мадиевой при участии сотрудников факультета филологии НИУ ВШЭ (Москва).
Для корпуса была адаптирована поисковая система Восточноармянского национального корпуса (EANC).
powered by Corpus Technologies
Точный
Неточный


форма
лемма
перевод
1
грамматика и части речи
Дополнительно
Расстояние до следующего слова:
от 1 до 1 (в словах)
форма
лемма
перевод
2
грамматика и части речи
Дополнительно
Расстояние: дополнительно
Искать
Очистить

Кроме того, для полноты данных была составлена таблица метаинформации, включающая в себя все выходные данные.

Нужно отметить, что Алматинский корпус

- в отличие от многих языковых корпусов, которые обладают только частеречной разметкой (а иногда не имеют грамматической разметки вообще), обладает полной морфологической разметкой;

- в отличие от большинства корпусов, содержит переводы слов на другой язык (русский), что облегчает работу пользователям, для которых казахский язык не является родным. Корпус обладает интерфейсами на трех языках. Например, в Национальном корпусе русского языка нет переводов слов на английский, а английский интерфейс обладает урезанными возможностями по сравнению с русским;

- обладает бесплатным общедоступным поисковым интерфейсом с мощным функционалом, что характерно для большинства корпусов, созданных за последние годы в рамках российской школы корпусной лингвистики и реже встречается в корпусах, создаваемых на Западе;

- в отличие от большинства корпусов малых языков России, созданных по схожей технологии в последние 5 лет, является относительно хорошо сбалансированным и содержит большое количество текстов, относящихся к художественной литературе.

Планируется, что в корпусе до конца года будет 2 млн. словоупотреблений. Сейчас в обработке находится 111 тыс.

Тексты корпуса, в первую очередь, предназначены для поддержки работы лингвистов, лексикографов, переводчиков, литературоведов, специалистов в области компьютерных исследований, организации образовательной среды в целях изучения и исследования казахского языка широким кругом как отечественных, так и зарубежных потребителей.

Алматинский корпус казахского языка способствует проведению фундаментально-прикладных исследований казахского языка на основе информационных технологий, внедрению их результатов в учебный процесс.

Корпус в связи с активным выходом Казахстана на мировую арену в последнее время приобретает активное значение и для преподавания и изучения казахского языка в качестве не только родного, но и иностранного. Немаловажное значение при этом, как уже говорилось, имеет то, что в отличие от других корпусов мира, казахский имеет перевод на русский и английский языки. В целях оптимизации преподавания родного и иностранного языков при составлении учебников имеется возможность наполнять их реальными примерами, что будет способствовать навыку развития естественных высказываний, поскольку у казахского языка нет такого широкого применения, в отличие от мировых и других более распространенных языков. Помимо этого существенным является то, что в настоящее время лексика казахского языка активно пополняется за счет английского и перевода уже имеющихся слов, ранее заимствованных из других языков.

По мере наполнения корпуса можно будет надеяться, что учебники и компьютерные обучающие программы (КОПР) будут ориентированы на корпус.

Кроме того, большое практическое значение корпус казахского языка имеет и при составлении лексикографических источников. Нужно учитывать, что в настоящее время лексикография казахского языка не изобилует источниками различных направлений. Этот фронт работы нуждается в максимальной доработке.

Корпус позволит осуществить формирование онлайнового электронного корпуса/подкорпусов текстов на казахском языке.

Безусловно, поскольку в настоящее время Алматинский корпус казахского языка составляет пилотную версию, он будет дополняться, обновляться как количественно, так и качественно, кроме того будет существенно улучшаться поисковая функциональность корпуса.

В перспективе для развития и усовершенствования Алматинского корпуса казахского языка предполагается следующее:

– лингвистически репрезентативный корпус;

– мощный поисковый аппарат для осуществления сложных лексико-морфологических запросов;

– удобный инструмент для самостоятельного изучения казахского языка, дающий для большинства словоформ лексико-морфологические разборы и русские/английские переводные эквиваленты;

– диахронически ориентированный корпус, покрывающий различные периоды истории современного казахского языка;

– диверсифицированный корпус, включающий разножанровые письменные и устные тексты разных типов;

– аннотированный корпус, снабженный грамматической и библиографической разметкой;

– корпус, находящийся в открытом доступе;

– электронная библиотека, включающая более 100 классических произведений казахской литературы.

Подводя итоги, можно сказать, что корпус казахского языка – это хранилище текстов, предназначенных для создания цельной информационной базы, дающей пользователю доступ пользователю к самому материалу как в его современном состоянии, так и в исторической перспективе.

Литература

1 Ельдесов Д. Язык без корпуса: возродится ли казахский язык? // http://www.altyn-orda.kz/dastan-eldesov-yazyk-bez-korpusa-vozroditsya-li-kazaxskij-yazyk/. – 2012. – 21 июня.

2 Сулейменова Э.Д. Языковая политика – фактор укрепления национально-государственной идентичности // http://dknews.kz/yazykovaya-politika-faktor-ukrepleniya-nacionalno-gosudarstvennojj-identichnosti/. – 2013. – 29 ноября.

3 Википедия // https://ru.wikipedia.org/wiki/.

4 Алматинский корпус казахского языка // http://web-corpora.net/KazakhCorpus/search/?interface_language=ru.

5 Корпус казахского языка // http://new.til.gov.kz/index.php/ru/the-corpus-of-kazakh-language.

6 Национальный корпус русского языка // http://www.ruscorpora.ru/.

7 Британский национальный корпус // http://www.natcorp.ox.ac.uk/.

References

1 Eldesov D. Yazyk bez corpusa: vozroditsya li kazakhskij yazyk? // http://www.altyn-orda.kz/dastan-eldesov-yazyk-bez-korpusa-vozroditsya-li-kazaxskij-yazyk/. – 2012. – 21 Yune.

2 Suleimenova E.D. Yazykovaya politika – factor ukrepleniya nacionalno-gosudarstvennoj identichnosti // http://dknews.kz/yazykovaya-politika-faktor-ukrepleniya-nacionalno-gosudarstvennojj-identichnosti/. – 2013. – 29 November.

3 Wikipedia // https://ru.wikipedia.org/wiki/.

4 Almaty Corpus of Kazakh // http://web-corpora.net/KazakhCorpus/search/?interface_language=ru.

5 The Corpus of Kazakh language // http://new.til.gov.kz/index.php/ru/the-corpus-of-kazakh-language.

6 Russian National Corpus // http://www.ruscorpora.ru/.

7 British National Corpus // http://www.natcorp.ox.ac.uk/.