Лингвистический корпус крымскотатарского языка

Radovan Garabík
Институт языкознания им. Л. Штура
Братислава, Словакия

Ленара Шакировна Кубединова
Таврический национальный университет им. В. И. Вернадского
Симферополь, Украина

Abstract. Corpus of Crimean Tatar language texts is aimed at the construction of a database of contemporary written Crimean Tatar language. The database consists mostly of journalistic texts, the corpus is freely accessible on the Internet, and searchable via a simple WWW form, using regular expressions for the queries. Some statistical results obtained from the corpus are described in the paper.

Абстракт: Корпус текстів кримськотатарської мови зосереджується на створення бази даних сучасної писаної кримськотатарської мови. База даних складена здебільшого із публіцистичних текстів. Корпус текстів є приступний на Інтернеті безплатно і можна використовувати його для вишукування посередництвом WWW інтерфейсу застосовуючи для пошуку регулярні вирази. В статті презентовані також статистичні результати здобуті у цього корпусу.

Введение

Корпусная лингвистика – современная, быстро развивающаяся область, возникшая вследствие растущих потребностей лингвистики во внедрении компьютерных технологий для работы с большими массивами языковых данных. На современном технологическом уровне лингвистику уже не удовлетворяют просто электронные библиотеки или полнотекстовые базы данных. Лингвисту нужны электронные корпуса, т. е. такие электронные коллекции текстов, которые отобраны исходя из некоторых принципов, специально подготовлены и размечены, и в которых с помощью специальных программ можно искать необходимые фрагменты текста по заданным параметрам. Внедрение корпусных методов радикально изменило общий научный ландшафт в лингвистике. Теперь ограничений на объем анализируемого материала и скорость поиска информации в нем по существу нет, а это означает, что в распоряжении исследователя оказываются колоссальные массивы текстов самого разного типа.

На сегодняшний день, каждый язык нуждается в создании корпуса, особенно, это касается малоизученных языков или языков, которые не располагают электронными технологиями для их изучения. Крымскотатарский язык не располагает ни своей электронной библиотекой, ни какой-либо полнотекстовой базой данных. Электронный корпус, т. е. компьютерная коллекция текстов, специально подобранных и подготовленных для научных исследований, может служить как их субституция. Лингвистический корпус крымскотатарского языка направлен на создание базы данных современного письменного языка.

Крымскотатарский язык

Крымскотатарский язык относится к тюркским языкам. Традиционно язык описывается как принадлежавший к кыпчакско-половецкой подгруппе кыпчакских языков, хотя средний диалект, на основе которого в 1928 году возник новый литературный язык, занимает положение между кыпчакскими и огузскими языками. Именно этот язык, начало кодификации которого было положено в 1920-е годы, и используется в общих чертах и по сей день.

Общая численность говорящих на крымскотатарском языке на территории бывшего СССР составляет приблизительно 400 тысяч человек, из них около 270 тысяч в Крыму. Диаспоры крымских татар находятся и в таких государствах, как Турция, США, Канада, Литва, Румыния, Болгария.

Орфография

До 1929 года крымскотатарский язык пользовался арабским алфавитом, с 1929 по 1938 латинским (так называемый Jaŋalif)[1], с 1938 — кириллицей. С

1 Языки мира: Тюркские языки. М.: Издательство «Индрик», 1997г. – 544с

1990-х годов осуществляется постепенный переход на латинизированный алфавит, утверждённый постановлением Верховного Совета Крыма в 1997 году. Надо подчеркнуть, что этот алфавит отличается от алфавита используемого в 1930-е годы. Издаваемые художественные тексты, журналы и газеты до сих пор используют преимущественно кириллицу. В корпусе решено в настоящее время предоставлять тексты на кириллице, с возможностью автоматической конверсии источников на латинском алфавите.

А/а Б/б В/в Г/г Гъ/гъ Д/д Е/е Ё/ё Ж/ж З/з И/и Й/й К/к Къ/къ Л/л М/м Н/н Нъ/нъ О/о П/п Р/р С/с Т/т У/у Ф/ф Х/х Ц/ц Ч/ч Дж/дж Ш/ш Щ/щ Ъ/ъ Ы/ы Ь/ь Э/э Ю/ю Я/я

Таб. 1: Крымскотатарская кириллица. гъ, къ, нъ *и* дж *считаются самостоятельными буквами.*

A/a Â/â B/b C/c Ç/ç D/d E/e F/f G/g Ğ/ğ H/h I/ı İ/i J/j K/k L/l M/m N/n Ñ/ñ O/o Ö/ö P/p Q/q R/r S/s Ş/ş T/t U/u Ü/ü V/v Y/y Z/z

Таб. 2: Крымскотатарская латиница

Лингвистическая обработка текстов

В предоставленном корпусе отсутствует сложный лингвистический анализ текста, кроме элементарной токенизации и определения пределов предложений.

Разработка средств автоматической лингвистической обработки для крымскотатарского языка находится на начальном этапе. Для корпуса важно существование морфологического анализатора и лемматизатора, который позволяет использовать корпус для сложных запросов, включающих разные грамматические категории, для поиска основных форм слов и для статистического анализа упомянутых категорий. На самом деле, агглютинативный характер крымскотатарского языка позволяет заместить

отсутствие поиска по леммам использованием регулярных выражений. Так например для того чтобы найти все формы слова *бала (ребенок)*, надо ввести регулярное выражение *бала.**, где . заменяет любой символ, и * соответствует нулю или более копий предыдущего (любого) символа, в результате чего такой запрос вернёт все формы (т. е. падежи) слова *бала*.

Число		Единственное	Множественное
Лицо	Падеж		
I	Имен. Род. Д-нап. Вин. Местн. Исход.	балам баламнынъ баламгъа баламны баламда баламдан баламнен	баламыз баламызнынъ баламызгъа баламызны баламызда баламыздан баламызнен
II	Имен. Род. Д-нап. Вин. Местн. Исход.	баланъ баланънынъ баланъгъа баланъны баланъда баланъдан баланънен	баланъыз баланъызнынъ баланъызгъа баланъызны баланъызда баланъыздан баланъызнен
III	Имен. Род. Д-нап. Вин. Местн. Исход.	баласы баласынынъ балагъа баласыны балада баласындан баласынен	баласы баласынынъ балагъа баласыны балада баласындан баласынен

Таб. 3: Парадигма слова бала

В будущем, перспективным может оказаться использование системы морфологического анализатора, которую описывает Altıntaş Kemal[2].

2 Altıntaş, Kemal: *A Morphological Analyser for Crimean Tatar*. In: Proceedings of TAINN 2001, June 2001, North Cyprus.

Структура корпуса

Тексты в корпусе следуют принципам, описанным в статье Радована Гарабика[3]. Тексты входят в четырёхуровневую иерархию – *архив*, *банк*, *корпусоид* и *дата*. В архиве данные сохраняются в оригинальном виде и формате. В банке тексты переведены в общий XML формат, содержавший чистые тексты с информацией об абзацах. На этом уровне тексты аннотированы простой библиографической разметкой, которая состоит из названия источника текста, жанра и (приблизительной) даты издания. Одному документу в архиве (например, одному году газеты) соответствует большое количество единиц в банке. Затем следует корпусоид, в котором тексты находятся в формате XCES[4], разделенные на слова и предложения. Структура корпусоида соответствует структуре банка. Наконец тексты поступают в секцию «дата», в цифровом формате корпусного менеджера.

Для поиска в корпусе используется система Manatee/Bonito, которая состоит из сервера (Manatee) и клиента (Bonito)[5], с самостоятельным пользовательским WWW интерфейсом. Сервер позволяет простой поиск одного слова, или фразы (нескольких слов в очередном порядке), или произвольных регулярных выражений. Manatee дальше позволяет проводить разные статистические исследования в рамках колокаций, как MI-score, T-score, и также частотный анализ разных форм найденных единиц (доступ к статистическому анализу

3 Garabík, Radovan (2005): *Corpus Construction Tools*. In: Труды международной конференции MegaLing'2005. Прикладная лингвистика в поиске новых путей. Ed. В. П. Захаров С. С. Дикарева. С.-Петербург: Издательство «Осипов» 2005, с. 26 – 32.

4 Ide, N., Bonhome, P., Romary, L. (2000): *XCES: An XML-based Encoding Standard for Linguistic Corpora*. In: Proceedings of the Second International Language Resources and Evaluation conference. Paris: European Language Resources Association

5 Rychlý, Pavel (2000): PhD Thesis: *Korpusové manažery a jejich efektivní implementace*, Faculty of Informatics, Masaryk University, Brno, Czech Republic.

возможен только с помощью клиента Bonito, но планируется присоединение статистических функций также к пользовательскому интерфейсу).

Пользовательский интерфейс

Цель корпуса – предоставить возможность лингвистического исследования, а также служить как база текстов, для лингвистов, писателей и всех интересующихся крымскотатарским языком. Корпус предоставлен бесплатно, для всех пользователей без необходимости регистрирования. Для улучшения возможности доступа, корпусом можно пользоваться не только с помощью самостоятельного клиента *Bonito*, но и путём веб-интерфейса на страничке корпуса (http://korpus.juls.savba.sk/QIRIM) с помощью любого браузера. Веб-интерфейс написан в языке программирования Python с помощью системы Karrigell[6]. Интерфейс на крымскотатарском, английском, русском и словацком языках. В интерфейс можно ввести либо самостоятельное слово, либо фразу – постепенность слов (разделенных пробелами), где каждое слово может быть или в полном виде, или с шаблонами регулярных выражений. Надо заметить, что каждая из графем *къ*, *гъ*, *нъ* и *дж* в регулярных выражениях считается двумя самостоятельными символами.

6 http://karrigell.sourceforge.net

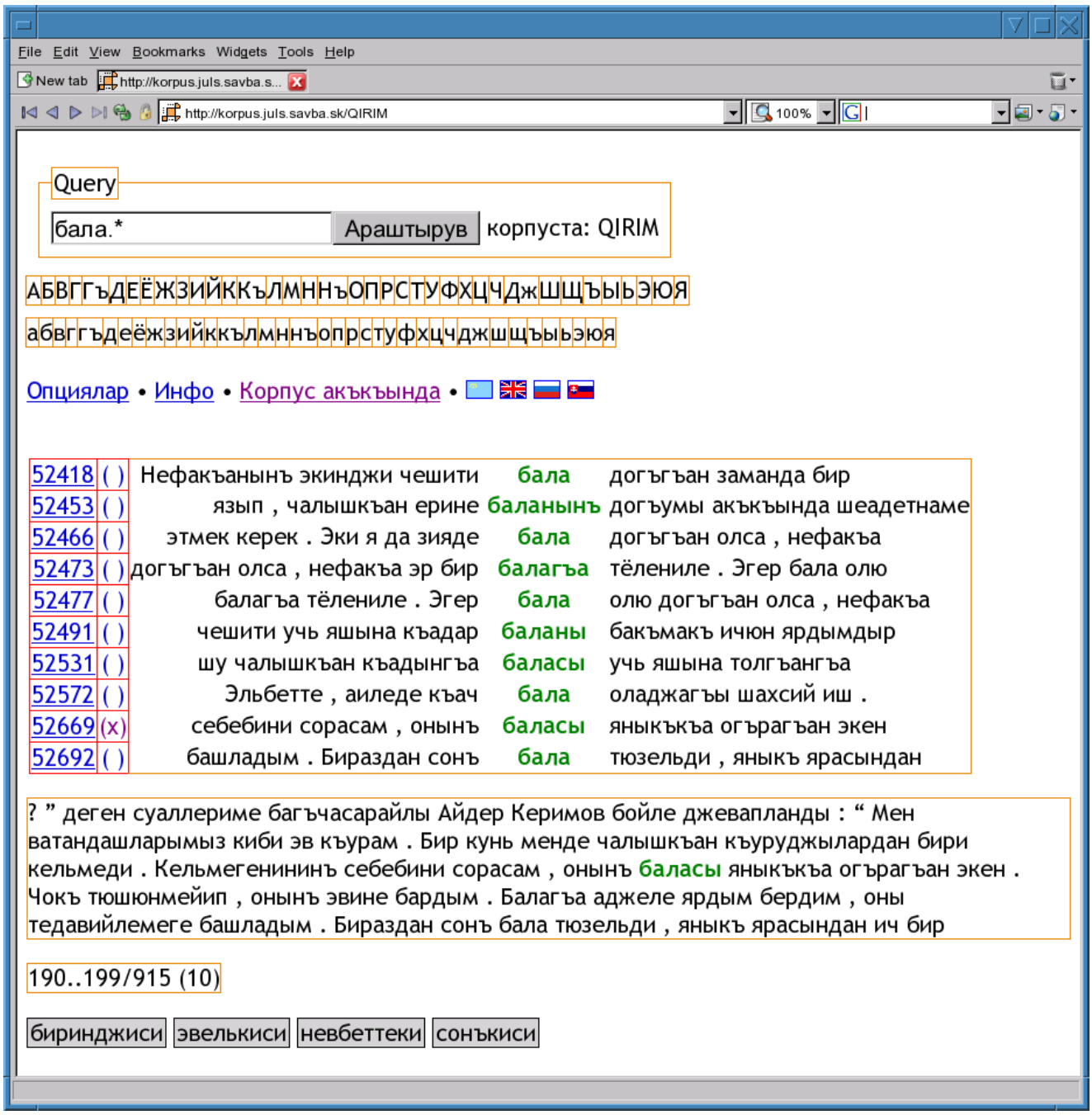

Рис. 1: Пользовательский интерфейс корпуса. Запрос регулярного выражения бала., т. е. все слова начинающиеся с* бала-. *Всех результатов 915, изображены с 190-ого до 199-ого, один с полным контекстом (внизу).*

Статистические исследования

В настоящее время корпус содержит 1988330 символов (букв), 371536 токенов (включая пунктуацию), что составляет 51965 разных словоформ. Из них 287804 «истинных» слов, без пунктуации и цифр (77.5%). Хотя по количеству

корпус принадлежит к малым корпусам, это уже даст возможность получить некоторые интересные статистические данные. В корпусе 24110 предложений, средняя длина предложения 15.4 токенов (т. е. слов, включая пунктуацию). Средняя длина слова (не включая пунктуацию и цифры) 6.58 символов (букв).

Самые длинные словоформы, находящиеся в корпусе:

радиоэшиттирювлеримизнинъ – наших радиопередач, где

радио – корень,

эшит- - корень (эшиттирмек),

* юв-* - аффикс отглагольного словообразования имён существительных

-лер - аффикс множественного числа,

-имиз - аффикс принадлежности настоящего времени 1 лица множественного числа,

-нинъ - аффикс родительного падежа;

укъукъкъорчалайыджыларгъа – правозащитникам, где

укъукъ – корень,

къорчалай- – корень

-ыджы- - аффикс отглагольного словообразования имён существительных,

-лар- аффикс множественного числа,

-гъа – аффикс дательно-направительного падежа;

словоформа	**количество**	**перевод**
ве	5598	и
бир	3113	один; одинаковый; некоторый
бу	1900	этот, эта, это
ичюн	1873	для; чтобы; из-за
де	1853	тоже, так же; и, да
да	1770	тоже, так же, ведь, хотя
эди	1390	инф. вспом. глаг.
Бу	1365	Этот, эта, это
сонъ	1172	потом, после
озь	1058	сам; свой; собственный

Таб. 4: Десять самых частотных слов в корпусе

Следующая работа и заключение

Основным параметром корпуса является, прежде всего, количество текстов. Поэтому мы предполагаем добавление новых текстов в корпус, в том числе и текстов написанных с использованием нового латинского алфавита. В связи с этим, необходимо разработать способ автоматической конверсии текстов между кириллицей и латиницей. С помощью программы, описанной в упомянутой статье[2], станет возможен автоматический морфологический анализ текста, что даст новые возможности поиска и статистического анализа текстов в корпусе. В настоящее время корпус является полезным источником текстов современного крымскотатарского языка с возможностью поиска лингвистических данных.