Ученые записки Крымского федерального университета имени В. И. Вернадского.
Серия «Филологические науки». Том 2 (68), № 3. 2016 г. С. 380–384.

УДК 81’33

МОРФОЛОГИЧЕСКАЯ РАЗМЕТКА КРЫМСКОТАТАРСКОГО ЭЛЕКТРОННОГО КОРПУСА (НА ОПЫТЕ ТАТАРСКОГО)

Кубединова Л. Ш.
Институт иностранной филологии, Таврическая академия, КФУ им. В. И. Вернадского

Гатиатуллин А. Р.
НИИ «Прикладная семиотика», Академия наук Республики Татарстан

В статье рассматривается первый этап системы морфологической разметки на уровне словоформы в крымскотатарском электронном корпусе. Представлены результаты сравнительного анализа аффиксальных морфем татарского и крымскотатарского языков.

Ключевые слова: электронный корпус, крымскотатарский язык, морфологическая разметка.

ВВЕДЕНИЕ

Лингвистический корпус крымскотатарского языка является на сегодняшний день первым и пока единственным корпусом национального языка крымских татар. Работа над ним совместно со старшим научным сотрудником Института языкознания им. Л. И. Штура Радованом Гарабиком началась еще в 2006 году [1]. Корпус пополняется в основном за счет публицистических текстов двух и единственных выпускаемых газет на крымскотатарском языке – «Янъы дюнья» и «Къырым». Первоначально работа велась только над текстами на кириллице. Впоследствии был создан небольшой подкорпус текстов на латинской графике, а также корпус крымскотатарской Википедии [2].

На начальном этапе в данном корпусе отсутствовал сложный лингвистический анализ текста, кроме элементарной токенизации и определения пределов предложений. Также отсутствие поиска по леммам, используя агглютинативный характер крымскотатарского языка, было замещено 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*.

В 2014 году совместно с научными сотрудниками научно-исследовательского института «Прикладная семиотика», Академии наук Татарстана началась работа по созданию морфологической разметки крымскотатарского электронного корпуса [3, 4].

ОПИСАНИЕ СИСТЕМЫ РАЗМЕТКИ

В электронном корпусе крымскотатарского языка идет работа по созданию различных систем разметок. Создание системы автоматического анализа морфологии

крымскотатарского языка является насущной и актуальной задачей. Она необходима для того, чтобы осуществлять поиск нужных пользователю слов.

Так, морфологическая разметка национального корпуса русского языка содержит информацию о морфологических формах и значениях (часть речи, род, число, падеж, наклонение и т. д.). В тюркских языках морфологическая разметка немного отличается. Там нет родов и такого строгого деления по частям речи, поэтому в предложенной нами разметке основам в результате анализа приписываются не части речи, а их морфологические категории в зависимости от присоединяемых аффиксальных морфем, форм слова, конструкций и т. д.

Система морфологической разметки в крымскотатарском электронном корпусе представляет собой разметку двух уровней:

– морфологическая разметка на уровне словоформы;

– морфологическая разметка на уровне аналитических форм.

В настоящее время реализован первый этап создания системы морфологической разметки на уровне словоформ и ведется работа по реализации системы разметки аналитических форм. Разметка системы аналитических форм предполагает подготовку базы данных со служебными словами крымскотатарского языка, которые используются для образования аналитических форм, а также системы тэгов для обозначения элементов этой базы данных.

Работа первого этапа по созданию системы морфологической разметки состояла из следующих этапов:

– сравнительный анализ системы разметок (тэгов) для других тюркских языков, в частности татарского и турецкого;

– разработка системы морфологических тэгов;

– разработка системы морфотактических правил крымскотатарского языка;

– подготовка словаря крымскотатарских основ с морфологической разметкой;

– заполнение информации в программный комплекс для многофункциональной модели тюркской морфемы.

Морфологическая разметка электронного корпуса основана на поморфемном разбиении крымскотатарской словоформы.

Например:

‹савускъанларнынъ›

1) N(савускъан) + PL (-лАр) + GEN (-нЫнъ)

‹Ава›

1) V(ав) + PRES_1 (-Й)

2) N*(ава)*

В примере показаны исходная словоформа и варианты результатов морфологического анализа. Результаты вариантов анализа имеют следующую структуру:

– корневая морфема с указанием морфологического типа;

– аффиксальные морфемы с указанием морфологической категории.

В нашей системе разметки используется 5 морфологических типов для корневых морфем: N, A, V, D, S. Количество морфологических типов не совпадает с частями речи, так как типы определяются наборами присоединяемых аффиксальных морфем.

Для обозначения аффиксальных морфем с морфологическими категориями приняли следующие определения.

Морфема в лингвистике определяется как минимальная значащая часть слова, совокупность морфов (алломорфов), имеющих одинаковое значение и ряд других общих признаков. В нашей системе разметок мы считаем одной морфемой те варианты, которые совпадают по правилам чередования и синтактике, но различаются по выражаемым значениям (т. е. многозначная морфема).

Обозначения морфологических категорий морфем необходимы для того, чтобы отличать омонимичные морфемы, выражающие разные морфологические категории и имеющие разные правила следования в словоформе. Для обозначения морфологических категорий использована система обозначений, описанная в работах И. А. Мельчука [5] и В. А. Плунгяна [6].

Например, в предыдущем примере использованы следующие обозначения грамматических категорий:

PL – PLURAL;
GEN – GENITIVE;
PRES_1 – PRESENT.

Несколько вариантов морфологической разметки связаны с тем, что морфологический анализатор может выдавать несколько вариантов морфологического анализа, как это показано в примере выше для словоформы *ава*. На данном этапе морфологической разметки корпуса указываются все варианты анализа без снятия морфологической неоднозначности. В дальнейшем планируется реализовать снятие морфологической неоднозначности с использованием механизмов контекстного снятия многозначности. Однако контекстно также не всегда возможно снять многозначность, поэтому планируется реализовать технологии ручного снятия многозначности экспертами.

Для составления системы аффиксальных морфем и морфологических категорий был проведен сравнительный анализ аффиксальных морфем татарского и крымскотатарского языков. Этот анализ показал, что среди крымскотатарских морфем есть целый ряд морфем, которые отсутствуют в татарском языке. Список этих морфем приведен в таблице 1.

Таблица 1

	Морфема	**Алломорфы**
1	-нен	-нен
2	-джА	-джа, -дже, -ча, -че
3	-мАктА	-макъта, -мекте
4	-Г[ъ]АйдЫ	-гъайды, -гейди, -къайды, -кейди

5	-мА+Й+Ып	-майып, -мейип
6	-мАлЫ	-малы, -мели
7	-АрАк[ъ]	-аракъ, - ерек, - яракъ
8	-мАдАн	-мадан, -меден
9	-ГЪАн+джА[къ]	-гъандже(къ), -гендже(к), -къандже(къ), -кендже(к)

Для всех этих морфем подготовлены свои тэги и морфотактические правила построения словоформ крымскотатарского языка, которые были использованы в морфологическом анализаторе.

Например:

-мАлЫ – DEB – Debitive
-Г[ъ]АйдЫ – OPT – Optative
-мАдАн – ADVV_NEG_2

ВЫВОДЫ

Таким образом, подготовлена система морфологической крымскотатарской словоформы и продолжается работа по созданию морфологической разметки аналитических форм. При создании следующих версий системы разметок крымскотатарского электронного корпуса планируется использование теоретических и практических работ, созданных крымскотатарскими лингвистами в области морфологии, синтаксиса и семантики крымскотатарского языка. В частности, это работы по синтаксису крымскотатарского языка Л. С. Селендили.

В виде проблемных моментов отметим: небольшой объем словаря крымскотатарских основ, который необходимо расширить, а также разное написание одних и тех же слов у разных крымскотатарских авторов.

Список литературы

1. Кубединова Л. Ш., Радован Гарабик Лингвистический корпус крымскотатарского языка // Прикладна лінгвістика та лінгвістични технології: MegaLing-2006:Зб. наук. пр. / НАН України. Укр. мовн.-інформ. фонд, Таврійськ. нац. ун-т ім. В.І. Вернадського; за ред. В. А. Широкова. – К.: Довіра, 2007. – С. 83–89.

2. Kubedinova Lenara. Corpus Linguistics: Studies in Crimean Tatar Language / Kubedinova Lenara, Radovan Garabik // TURKLANG'14 International Conference on Turkic Language Processing, 6–7 November 2014 – http://turklang.itu.edu.tr/invited-speakers.htm

3. Kubedinova Lenara, Gatiatullin Ayrat. Morphollogical tagging of Crimean Tatar electronic corpus / Kubedinova Lenara, Gatiatullin Ayrat // Proceedings of the International Conference «Turkic Languages Processing: Turklang-2015». – Kazan: Academy of Sciences of the Republic of Tatarstan Press, 2015. – 331–337 c.

4. Кубединова Л. Ш., Гатиатуллин А. Р. О реализации системы морфологической разметки крымскотатарского электронного корпуса // Труды Международной конференции по компьютерной и когнитивной лингвистике TEL-2016. – Казань: Изд-во Казан. ун-та, 2016. – С. 90–94.

5. Мельчук И. А. Курс общей морфологии. Т. IV. / Пер. с фр. Е. Н. Саввиной под общ. ред. Н. В. Перцова. – М., Вена: Языки славянской культуры: Венский славистический альманах, 2001. – 584 с.

6. Плунгян В. А. Общая морфология: Введение в проблематику: Учебное пособие. М.: Эдиториал УРСС, 2000. – 384 с.

Кубедінова Л. Ш., Гатіатуллін А. Р. Морфологічна розмітка кримськотатарського електронного корпусу (на досвід татарського) / Л. Ш. Кубедінова, А. Р. Гатіатуллін // Вчені записки Кримського федерального університету імені В. І. Вернадського. – 2016. Серія: Філологічні науки. – Т. 2 (68), № 3. – С. 380–384.

У статті розглядається перший етап системи морфологічної розмітки на рівні словоформи в кримськотатарському електронному корпусі. Представлені результати порівняльного аналізу афіксальних морфем татарської і кримськотатарської мов.

Ключові слова: електронний корпус, кримськотатарська мова, морфологічна розмітка.

Kubedinova L. Sh., Gatiatullin A. R. Morphological tagging of crimean tatar electronic corpus (by experiment on tatar language) / L. Sh. Kubedinova, A. R. Gatiatullin // Scientific Notes of Crimean Federal V. I. Vernadsky University. – Series: Philological Science. – 2016. – Vol. 2 (68), No. 3. – P. 380–384.

Nowadays many text electronic corpuses are created for many languages of Turkic group. Such corpuses already exist for Turkish, Tatar, Kazakh, Bashkir, Tuvinian and other Turkic languages. All authors of these corpuses are faced the same problems and start heading the same way of creating their own systems of corpus annotation. Although structure similarity of Turkic languages allows to create a common base of computer and program models for processing texts in Turkic languages.

The work on the Linguistic corpus of Crimean Tatar language started in 2006 jointly with a senior researcher of L. I. Stura Institute of linguistics Radovan Garabik. The corpus is mostly supplemented with texts from the only two Crimean Tatar newspapers «Yanı dünya» and «Kırım». At the beginning there were used only the Cyrillic texts in Crimean Tatar language. Afterwards a subcorpus in Latin script and the corpus of Crimean Tatar Wikipedia were created.

The work on morphological tagging of Crimean Tatar electronic corpus started in 2014 in cooperation with researches of Research Institute of Applied semiotics of Tatarstan Academy of Sciences.

In this article the system of morphological Crimean Tatar wordforms is suggested and the work on the creating of morphological tagging of analytical forms is proceeding. This system is developed on the basis of tags which are used for annotation of electronic corpus of Tatar language «Тугантел» («Mother tongue»). The results of the comparative analyses of affix morphemes of Tatar and Crimean Tatar languages were represented. While creating next tagging systems of Crimean Tatar electronic corpus it is planned to use theoretical and practical works of Crimean Tatar linguists in fields of morphology, syntax and semantics of Crimean Tatar language.

Keywords: electronic corpus, Crimean Tatar language, morphological tagging.