Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка

Б. В. Орехов[1], Е. А. Слободян[2]
[1]Башкирский государственный педагогический университет им. М. Акмуллы, Уфа; Университет Осло, Норвегия;
[2]Башкирский государственный педагогический университет им. М. Акмуллы, Уфа

We explore the problems of automatic analysis of the word morphologic structure concerning the Turkic languages and Bashkir. We report about a special program for automatic morphologic analysis of the Bashkir word and some of its prospects for science and technology.

Важным, едва ли не центрообразующим звеном цепи автоматической обработки текста на естественном языке является технология нахождения основы слова (стемминг), родственный ей по целям алгоритм (лемматизация), позволяющий определить, что некоторая цепь словоформ составляет одно «словоизменительное гнездо» (имеет одну лемму). Конечным продуктом, способным на эти операции, является программа (парсер), в автоматическом режиме осуществляющая морфологический разбор слова. Это и основа качественного информационного поиска (большие поисковые системы в Интернете стремятся учитывать при поиске морфологию языка запроса-выдачи), и технологическая база для создания лингвистического корпуса языка.

Как ни странно, для башкирского языка такой технологии реализовано не было, по крайней мере, в открытых источниках информацию о ней найти не удалось. Возможно, это связано с пока ещё низкими потребностями в обработке электронных текстов на башкирском языке: так, число сайтов на башкирском языке в Интернете не превышает нескольких десятков. Ясно, что это число будет со временем расти, и автоматическая обра-

ботка текстов на башкирском языке — перспективная отрасль компьютерной лингвистики, не говоря уже о назревшей необходимости создания корпуса башкирского языка. Существующий на данный момент Машинный фонд башкирского языка нельзя назвать полноценным корпусом, потому что в нём отсутствует возможность работы с репрезентативным и сбалансированным по жанрам количеством текстов на данном языке.

Образовавшаяся лакуна тем более заметна, что построение парсера для агглютинативных языков — задача, в принципе, не самая сложная. Словоизменительный принцип, подразумевающий регулярный порядок присоединения аффиксов к основе, низкий процент грамматической омонимии, отсутствие инфиксов и прочих подобных случаев, затрудняющих реализацию автоматического распознавания грамматической формы слова, крайне облегчают практическую реализацию программного средства для распознавания формы слова. Лингвисты-любители даже заподозрили тюркские языки в искусственном происхождении, имея в виду их «нечеловеческий» математический конструктивный принцип [Абдульманова]. Это, конечно, можно расценивать только как печальный курьёз, но курьёз по-своему показательный. Он демонстративен не только в смысле факта лингвистической безграмотности, к сожалению, присущей представителям других научных дисциплин, но и как символический знак высокой алгоритмизованности, заложенной в порождающую модель тюркских языков.

Специалистами уже разрабатывались алгоритмы стемминга для тюркских языков, в основном, турецкого. Так, B. Taner Dinçer и Bahar Karaoğlan в 2003 году представили работу под названием «Stemming in Agglutinative Languages: A Probablistic Stemmer for Turkish» [Taner Dinçer]. Авторы утверждают, что разработанный ими алгоритм успешно применяется в 95,8% случаев.

Однако стемминг — это локальная проблема, имеющая значение только для прагматических задач информационного поиска. Более полное отражение дискутируемый вопрос находит во взаимообратимых алгоритмах парсинга и формопорождения.

Формальное описание алгоритмов формопорождения при башкирском словоизменении сделано З. А. Сиразитдиновым [Сиразитдинов] в рамках задачи построения программы для

проверки орфографии башкирского языка (спелл-чекера). Фактически процесс порождения словоформ от заданной леммы при работе спелл-чекера и процесс морфологического разбора являются (с незначительными отличиями) проявлениями работы в разных направлениях одного и того же алгоритма, более или менее моделирующего естественный языковой процесс построения и распознания словоформ говорящим и слушающим.

Благодаря этому описанию обнаруживаются необходимые и достаточные условия для создания программного решения. Во-первых, как и в случаях с аналогичными разработками для других языков, в основе программы (и спелл-чекера, и парсера) должен лежать грамматический словарь, то есть набор слов (лемм), разбитый на словоизменительные типы. В случае с башкирским языком такой словарь должен включать в себя типы существительных, прилагательных, глаголов, наречий и неизменяемых слов.

Широко известно, что большим преимуществом русского языка перед многими другими в смысле создания программ для автоматической обработки текстов на естественном языке в своё время стало наличие грамматического словаря русского языка, созданного А. А. Зализняком [Зализняк]. Это фундаментальное издание включало описание словоизменения для абсолютного большинства русских слов. У башкирского языка также имеется подобное преимущество. В 1994 году в свет вышел Грамматический словарь башкирского языка М. Х. Ахтямова [Ахтямов]. Однако внимательное знакомство с этим лексикографическим трудом, к сожалению, обнаруживает, что в силу понятных и естественных причин (ориентация на использование в практической деятельности человека-специалиста, носителя языка, а не в работе автомата) использование грамматического словаря в работе программы невозможно. Одной из обусловливающих такое обстоятельство тонкостей является принципиальное (и существенное для грамматики) разделение лексического фонда башкирского языка на собственно башкирские слова и русские заимствования, причём основы последних присоединяют определённый тип аффиксов, подчиняясь иным правилам, нежели те, которые действуют для собственно башкирских слов. Это значимое разделение в словаре М. Х. Ахтямова не отражено, вероятно, потому, что для носителя языка проис-

хождение слова очевидно, но, к сожалению, не очевидно для автоматической программы. Кроме того, лишним для той же автоматической программы можно признать простое деление внутри неизменяемых слов (как самостоятельных, так и служебных). Это не значит, что в контексте автоматического разбора слова деление, например, на послелоги и союзы бессмысленно. Наоборот, при существительном послелог, обусловливающий выбор формы некоторого падежа, мог бы быть дополнительным критерием определения грамматической формы субстантива в тексте, помогающим избежать ошибок в разборе слова.

В настоящее время программное решение для морфологического разбора башкирского слова (с ориентацией на разбор не отдельных слов, а целых текстов) нами реализовано на платформе языка Perl и сейчас находится в процессе тестирования его эффективности[1]. Автомат последовательно вычленяет аффиксы и находит основу, то есть выполняет функции стеммера, одновременно с этим на основе грамматического значения найденных аффиксов характеризуя морфологическую форму слова, то есть является полноценным парсером.

По мере апробации программы на реальных текстах сбалансированной жанровой принадлежности (как художественной, так и нехудожественной литературы) следует ожидать расширения словарной части программы, той, которая содержит необходимый для парсинга список основ. В настоящее время можно признать особенно ущербной ту часть списка лемм, которая включает в себя собственные имена — не только антропонимы, но и топонимы. Эта проблема будет постепенно решаться в процессе совершенствования программы по принципу «обучения с учителем».

Создание парсера башкирского языка открывает перед лингвистами и специалистами по информационному поиску несколько перспектив.

Первой, фундаментальной, задачей нам видится создание корпуса башкирского языка, соответствующего современным стандартам, предъявляемым к таким разработкам корпусной

[1] Программа (под наименованием «Морфологический анализатор башкирского языка») вошла в Единый реестр инновационных проектов Республики Башкортостан под номером 170.

лингвистикой. Преимущества лингвистов, обладающих таким исследовательским инструментом по сравнению с теми, кто им не обладают, представляются нам очевидными.

Предсказуемые лингвистические результаты такой деятельности следующие. Общеизвестно, что описания в грамматиках, тяготеющие к кодификации и унификации, зачастую представляют реальную языковую ситуацию излишне идеализированно, не учитывают естественно существующие в языке нелинейные явления. Обработка башкирских текстов парсером в первую очередь позволит на обширном текстовом материале выявить частотность отклонений тех или иных явлений от академических грамматических описаний, то есть получить более полное представление о функционировании языка.

Второй, более практической задачей стала бы разработка специализированной поисковой системы по башкирскому сектору Интернета с учётом морфологии башкирского языка. Такая поисковая система могла бы благотворно повлиять и на расширение башкирского Интернета.

Список литературы

Абдульманова — Абдульманова Э. Татарские ученые, возможно, разгадали код татарского языка [Электронный ресурс] // Tatarlar.ru: [сайт]. URL: http://www.tatarlar.ru/news.php?lng=ru&pg=12026 (дата обращения: 09.09.2010).

Ахтямов — Ахтямов М. Х. Грамматический словарь башкирского языка (на баш. яз.). Уфа, 1994.

Сиразитдинов — Сиразитдинов З. А. Алгоритмическая грамматика словоизменения башкирского языка: [сайт]. URL: http://mfbl.ru/bashdb/algram/algram.htm (дата обращения: 19.09.2010).

Зализняк, 1977 — Зализняк А. А. Грамматический словарь русского языка. М., 1977.

Taner Dinçer B. Bahar Karaoğlan Stemming in Agglutinative Languages: A Probablistic Stemmer for Turkish // Computer and Information Sciences. ISCIS 2003. Berlin: Springer Berlin; Heidelberg, 2003. P. 244–251.