УДК 004.658.6

Л.Е.Урмазова (2 в.о., каф. ИУС), А.В.Шокало (асп., каф. ИУС)

РАЗРАБОТКА СИСТЕМЫ МОРФОЛОГИЧЕСКОГО АНАЛИЗА ПОЛЬСКОГО ЯЗЫКА

Системы автоматической переработки текста начали разрабатываться еще в середине XX века, однако, спустя десятилетия, исследователи выяснили, что область применения таких систем ограничена ввиду недопустимости отождествления возможностей компьютера с речемыслительными возможностями человека.

В настоящее время продолжается разработка различных алгоритмов анализа текста, но их использование сводится, в основном, к переводу технической документации, поисковым системам или системам автоматического реферирования.

При создании таких систем наиболее часто используется т.н. модульно-иерархическая организация, в которой за обработку текста на каждом уровне – лексическом, морфологическом, синтаксическом, семантическом и т.д. – отвечает отдельный модуль, а выходные данные k-го модуля поступают на вход k-1-го модуля.

Независимо от назначения системы, обязателен «нижний» модуль, отвечающий за морфологический анализ.

Обычно работа модуля обеспечивается обращениями к т.н. лингвистической базе данных (ЛБД), в которой обязательно присутствует таблица – автоматический словарь, включающая в себя поля словоформ и соответствующей им грамматической информации. В случае обработки синтетических языков вместо словоформ в таблицу заносятся т.н. машинные основы – неизменяемые части основ, а словарь дополняется отдельной таблицей – машинной морфологией, включающей флексии, относящиеся к машинным основам. Эти таблицы связываются по полю грамматической информации отношением «один-ко-многим».

Однако существуют синтетические языки, для которых перечисленных таблиц недостаточно или такая традиционная структура не может обеспечить адекватность выходных данных. К таким языкам относится польский – западно-славянский язык синтетического типа с разветвленной системой чередований в основе. Включение неизменяемых частей основы в автоматический словарь приведет к возможным ошибкам поиска. Например, включение неизменяемой части основы слова *miasto/mieście – город (пол.) – «mi-»* в автоматический словарь и ее последующий поиск выдаст несколько омонимичных основ, выбор из которых может затруднить работу следующего уровня системы, а поиск основы слова *sen/śnie – сон (пол.)* не даст результатов, потому что основа в этом случае окажется нулевой (пустой).

В качестве возможной реализации системы морфологического анализа польского текста предлагается использовать расширении базы данных за счет введения дополнительных сущностей – чередования в основе и номера формы. Тогда в автоматический словарь вносятся основы начальной формы слова, а чередования помещаются в отдельную таблицу. В эту же таблицу вносятся номера форм, где произошло чередование (например, для слова *miasto* в словарь заносится основа *miast*, а чередование *ast/eść* включается в таблицу чередований). Эти таблицы также связываются по полю лексико-грамматической информации. Парадигма слова с данным кодом лексико-грамматической информации (то есть окончания всех форм слова) вносятся в таблицу машинной морфологии. Для слова *miasto* это *–o, -a, -u, -em* и т.д. Номер формы (например, для существительных – от 1 до 14) вносится в таблицу машинной морфологии и таблицу чередований.

Алгоритм анализа входной словоформы выглядит следующим образом:

1. Делаются предположения о возможной границе между основой и окончанием словоформы. Для этого последовательно отсекаются символы словоформы, начиная с

начального. Полученные таким образом и найденные в базе флексии со значениями полей сопровождающей их грамматической информации запоминаются. Часть слова после отсечения каждой из найденных флексий считается возможной основой.

2. Проверяется наличие каждой из возможных основ в базе. Если ни одна из них не найдена, то делается предположение о наличии чередований в основе.
3. От каждой из проверяемых основ последовательно отсекаются символы, начиная с начального. Затем производится поиск каждой получившейся комбинации в таблице чередований.
4. Поле грамматической информации каждого из найденных чередований и поле номера их форм сравнивается с аналогичными полями, характеризующими флексию, при отсекании которой получилась соответствующая основа. Результатом является одно значение поля лексико-грамматической информации и номера формы, а также часть основы до чередования.
5. От основы отсекается чередование и присоединяется соответствующая ему часть основы до чередования. Результатом является основа начальной формы.
6. Выполняется поиск полученной основы начальной формы в базе. Если он результативен, анализ завершен. В противном случае слово считается отсутствующим в базе.

В настоящее время разработана база данных на СУБД MS SQL Server и пользовательский интерфейс, позволяющий пользователю ввести польское слово в любой форме и получить на выходе его начальную форму, грамматическую информацию о части речи, форме, в которой находится входное слово.