УДК 519.168

МОРФОЛОГИЧЕСКИЙ СТАНДАРТ НАЦИОНАЛЬНОГО КОРПУСА ЧУВАШСКОГО ЯЗЫКА

Желтов В.П.[1], Желтов П.В.[1], Губанов А.Р.[1]

[1]ФГБОУ ВПО «Чувашский государственный университет им. И.Н. Ульянова», Чебоксары, Россия (428015, Чебоксары, Московский пр-т, 15), e-mail: zheltov42@mail.ru

Создание Национальных корпусов языков, а также такое направление, как автоматическая обработка естественного языка, ставят задачу разработки морфологических стандартов представления данных. В связи с этим в статье рассмотрен морфологический формат, позволяющий создавать Национальный корпус чувашского языка и составляющий теоретическую основу морфологической аннотации чувашских текстов. Отмечены особенности чувашского языка, востребованные для грамматической аннотации: имеет развитую систему грамматически однозначных словоизменительных аффиксов, в нем отсутствуют различные парадигматические классы в парадигме того или иного одного типа, отсутствуют значительные чередования в основах, а также закономерная фонетическая обусловленность алломорфов и т.д. Используя представленную схему, можно описать все чувашские фонологические правила, которые будут отражать все фонологические явления в чувашском языке. В статье особое внимание уделено грамматическому словарю чувашского языка, который с программной точки зрения может представить базу данных и диалоговый интерпретатор запросов к базе данных.

Ключевые слова: национальный корпус языков, автоматическая обработка естественного языка, морфологический стандарт, чувашский язык, фонологические правила, грамматический словарь чувашского языка, база данных

MORPHOLOGICAL STANDARD OF THE NATIONAL CORPUS OF THE CHUVASH LANGUAGE

Zheltov V.P.[1], Zheltov P.V.[1], Gubanov A.R.[1]

[1]Federal state budget educational institution of higher professional education "Chuvash State University named after I.N. Ulyanov", Cheboksary, Russia (428015, Cheboksary, Moskovsky Prospect, 15), e-mail: zheltov42@mail.ru

Creating a National Corps languages and such direction as automatic processing of natural language, are aiming to develop standards for the presentation of morphological data. In this regard, the article discusses the morphological format, allowing you to create the National Corpus of the Chuvash language, and which constitutes the theoretical basis of morphological annotations Chuvash texts. The features of the Chuvash language, demand for grammatical annotation: has developed a system of unambiguous grammatical inflectional affixes, it lacks various paradigmatic classes in the paradigm of a single type, no significant alternation in the basics, as well as a natural phonetic allomorphs conditioning, etc. Using the representation scheme can describe all the Chuvash phonological rules that will reflect all the phonological phenomena in the Chuvash language. The article focuses on the grammar dictionary Chuvash language, which has a program point of view, may present a database and an interactive interpreter database queries.

Keywords: national body language, automatic natural language processing, morphological standard, the Chuvash language, phonological rules, grammar dictionary Chuvash language, the database

В современной прикладной лингвистике исследование естественных языков на эмпирически достоверном материале с использованием технологий автоматической обработки языковых данных представляет перспективное междисциплинарное направление. Эффективным средством решения лингвистических задач являются Национальные корпусы русских и тюркских языков, в том числе и Национальный корпус чувашского языка (НКЧЯ). Создание подобной системы для чувашского языка позволяет получать новые данные о структуре языка, о его лексическом составе и дает ценный материал для дальнейших исследований в лингвистических моделях и реализации технологий автоматической

обработки текстов (АОТ) на естественных языках, в частности тюркских.

Создание Национальных корпусов языков, а также такое направление, как автоматическая обработка естественного языка, ставят задачу разработки морфологических стандартов представления данных. К тому же корпуса, создаваемые на языках народов России, разные по объему, технологиям и иному, потому что неоднородны сам языковой материал и корпусная тематика. В связи с этим научно-практическим семинаром «Унификация систем грамматической разметки в корпусах тюркских языков (семинар UniTurk)» принята своевременная резолюция, где говорится, что «создание электронных лингвистических корпусов выдвигает перед разработчиками широкий спектр проблем и задач, успешное решение которых требует соединения результатов лингвистических исследований и современных компьютерных методов анализа языковых данных. Возможности корпуса во многом определяет система аннотации (разметки)». Участники научно-практического семинара «Унификация систем грамматической разметки в корпусах тюркских языков (семинар UniTurk)» указывают на то, что одной из важнейших задач тюркского языкознания является выработка такого стандарта представления лингвистической информации, который бы позволил организовать существующие и создающиеся корпуса тюркских языков в единое информационное пространство для широкого круга пользователей – специалистов-тюркологов, типологов и неспециалистов.

Как мы видим, актуальной задачей является разработка определенных стандартов представления данных: стандартизация форматов и стандартизация концепций. В частности, морфологический стандарт для систем автоматической обработки чувашских текстов обеспечивал бы единообразное представление информации; составлял бы теоретическую основу морфологической аннотации.

Отметим характеристики чувашского языка, востребованные для грамматической аннотации. Чувашский язык как язык агглютинативного типа имеет развитую систему грамматически однозначных словоизменительных аффиксов (отдельно взятый аффикс выражает один морфологический признак, в нем отсутствуют различные парадигматические классы в парадигме того или иного одного типа; отсутствуют столь значительные чередования в основах, а также закономерная фонетическая обусловленность алломорфов (границы морфем четкие: к основе присоединяются аффиксы с тем или иным значением, а если происходят фонемные изменения на границах морфем, то данные морфонологические изменения связаны с фонологическими законами чувашского языка)). Что касается автоматизации морфологической разметки чувашского текста, то следует отметить, что грамматические характеристики чувашских частей речи достаточно полно описаны в многочисленных трудах по чувашскому языкознанию.

При автоматическом анализе аффиксального состава словоформ в чувашском языке грамматические (морфологические) признаки распознаются относительно легко. Однако выявление и автоматическая разметка семантических (лексико-грамматических разрядов) различных частей речи не могут быть решены с учетом лишь морфологических данных. Поэтому на первом этапе создания Национального корпуса чувашского языка морфологическая разметка должна содержать, на наш взгляд, информацию о морфологических категориях, явно выраженных аффиксами.

В процессе автоматической разметки текстов следует учесть так называемые проблемы в регулярной морфологии чувашского языка и их возможности. С точки зрения отношения к системе языка соответствующие проблемные нарушения можно подразделить на 2 вида: внешние (несистемные) и внутренние (системные).

К внешним нарушениям морфологии чувашского относятся в первую очередь правила морфологического изменения неассимилированных заимствований и случаи, вызванные несовершенством современной чувашской орфографии. В условиях существующей чувашской орфографии и действующих принципов освоения заимствований возникает проблема автоматической обработки этих случаев. Необходимо решить, как обрабатывать заимствования, и найти способы описания возможных закономерностей их изменения.

Помимо указанных внешних факторов, существуют внутренние языковые особенности, которые являются системными и потенциально автоматически распознаваемыми. Подобные особенности есть следствие универсальности, полифункциональности, экономичности, присущих языковым элементам, что является их системным свойством и подчиняется определенным глубинным закономерностям.

Исходя из особенностей морфологии чувашского можно сказать, что трудности при автоматической разметке способны создать полифункциональные и омонимичные аффиксы, так называемые нулевые формы, особенности однородных членов (аффикс может присоединяться только к последнему члену группы), особенности изменения отдельных категорий слов (в частности, некоторых местоимений и послеложных слов).

Формализованное представление морфологических характеристик произвольного текста на чувашском языке в электронном машиночитаемом формате на первоначальном этапе создания корпуса следует ограничить в морфологической разметке информацией о грамматических категориях, явно выраженных аффиксами. Эффективность повысится путем введения фонологических и морфотактических правил.

Учитывая особенности морфотактики в чувашском языке, грамматические параметры именных слот можно подразделить, с одной стороны, на сложные и простые, а с другой —

на обязательные и факультативные. Что касается обязательности или факультативности параметра, то обязательный параметр должен быть приписан любой словоформе определенного морфологического класса (в частности, существительные обязательно стоят в каком-либо падеже, нет «беспадежных» форм существительных). Факультативный параметр, помимо какого-либо конкретного значения, может также принимать отрицательное значение (отсутствие признака — существительные в чувашском языке могут и не выражать значение принадлежности).

В морфотактической схеме слота «Имя существительное» участвуют, в частности, такие классы аффиксов, связанных с обязательными параметрами падежа (ЛФ – лексическая форма; ПФ – поверхностная форма):

N CASESI – класс падежных аффиксов:

ЛФ: пӳрт (дом) + *ĕн (н) ăн (н)* (CASE GEN)

ПФ: пӳртен

ЛФ: пӳрт + е а (CASE DIR)

ПФ: пӳрте (дому)

ЛФ: пӳрт + ***ре (те, че)*** + ***ра (та***) (CASE LOC)

ПФ: пӳртре (доме)

ЛФ: пӳрт + *рен (тен)* + *ран (тан)* (CASE ABI))

ПФ: пӳртрен (из дома)

ЛФ: пӳрт (дом) + *пе (пеле, пелен)* (CASE ABL)

ПФ: пӳртпеле (домом)

ЛФ: пӳрт + +*сĕр (сăр*) (CASE DER)

ПФ: пӳртсĕр (без дома)

ЛФ: пӳрт + *шĕн (**шăн***) (CASE CAUSE)

ПФ: пӳртшĕн (из-за дома)

В морфотактической схеме рассматриваемого слота участвуют также и другие классы аффиксов, связанных с факультативными параметрами.

Эффективность стандарта, как уже было сказано, также повысится путем введения фонологических правил, которые состоят из следующих компонентов:

1) связь-соответствие между лексическим и поверхностным символами, в частности соответствие м:н – çын+сем – çын+сен;

2) определяющий данное соответствие операторы (для их обозначения используются математические символы): а) => проявление соответствия только в данном контексте, но не всегда; б) <= проявление соответствия в данном контексте всегда, но не только в нем; в) <=> проявление соответствия в данном контексте всегда и только в нем.

Используя соответствующую схему, можно описать все чувашские фонологические правила, которые будут отражать все фонологические явления в чувашском языке: законы сингармонизма, гармонию согласных, а также случаи и исключения, возникающие при «осложнении» ЛФ, рассмотренных выше, смысловыми отношениями. Кстати, если внедрить функциональный подход и в фонологические правила, то это послужило бы базой для синтаксического анализатора (СА) для снабжения чувашских текстов детальной не только морфологической, но и предбазой синтаксической информации, когда даже современные существующие СА (в системах Dialing, ЭТАП и др.) эту базу создают с «нуля», имеем в виду отношения «субъект – действие», «объектные отношения» «темпоральные отношения», «каузальные отношения» и др. (если учесть то обстоятельство, что архитектура тюркских синтаксических анализаторов в корне отличаются от русских в связи с их морфотактикой, автоматной морфологией и их функциональностью, что минимизирует проблемы, связанные со структурой текста (синтаксиса)). Уместно отметить и то, что такой подход установил бы также «задел» для актуальной в современной лингвистике (и не только в лингвистке) системы автоматизации построении онтологии.

Модели лингвистической информации, семантико-грамматические аннотации лексем, построенные на основе типологических особенностей чувашского языка, можно представить в созданной лексикографической базе инверсионного, грамматического словаря – Обратного словаря чувашского языка [7], ибо практическая значимость словарей такого типа заключается в группировке слов по одинаковому концу. Для чувашского языка данный принцип особенно важен, так как аффиксы в нем располагаются справа от корня. Слова в инверсионном словаре в дальнейшем можно сгруппировать по морфологическому признаку (часть речи, наличие или отсутствие того или иного суффикса). В частности, анализ существующих Обратных словарей на практике позволил нам представить многообразие суффиксальных средств имен в чувашском языке, их продуктивность. В обратном словаре имеются массивы слов (более тысячи), которые имеют определенный суффикс. Следует обогатить будущий Грамматический словарь чувашского языка материалами новых словарей чувашского и русского языков, за счет чего расширится круг представленных в них характеристик. В перспективе обратный словарь с программной точки зрения можно представить как базу данных и диалоговый интерпретатор запросов к базе данных в разработанной нами ранее системе Java (в этом смысле имеются опережающие нас работы, связанные не только с чувашским языком, но и в сопоставлении с русским) [5].

Таким образом, Обратный словарь с перечисленными нами характеристиками, т.е. с наиболее полной информацией о грамматической характеристике лексики чувашского языка, может быть использован для квантитативного описания по широкому кругу

морфологических характеристик, а также значительно расширил бы информационную базу Национального корпуса чувашского языка.

Концептуальный подход к стандарту морфологической модели корпуса чувашского языка (что является основой его модели морфологической разметки чувашских текстов) позволяет создавать формализованное представление морфологических характеристик произвольного текста на чувашском языке в электронном формате. В связи с этим следует отметить, что в татарской компьютерной лингвистике на основе разработанного ими морфологического стандарта уже ведутся работы в области морфологической коррекции татарского текста [4]. На первоначальном этапе создания Национального корпуса чувашского языка обоснованным представляется ограничиться в морфологической разметке информацией о грамматических категориях, явно выраженных аффиксами. Эффективность же автоматической разметки можно повысить путем введения дополнительных фонологических и морфотактических правил для морфологического анализатора.

Публикация подготовлена в рамках поддержанного РГНФ научного проекта № 15-04-00532.

Список литературы

1. Апресян Ю.Д. Идеи и методы современной структурной лингвистики. М.: Наука, 1966. – 305 с.
2. Барахнин В.Б., Лукпанова Л.Х., Соловьев А.А. Алгоритм синтеза словоформ казахского языка с использованием флективных классов // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Дубна, Россия, 13–16 октября 2014 г. – С. 108–112.
3. Володин А.П., Храковский В.С. Типология морфологических классификаций глагола (на материале агглютинативных языков) // Типология грамматических категорий: Мещаниновские чтения. М.: Наука, 1975.
4. Гатиатуллин М.Р. Технология морфологической коррекции татарского текста // Труды Казанской школы по компьютерной и когнитивной лингвистике. Казань. 2000.
5. Димитриев А.П., Алексеев М.Б. База данных, применяемых в программе чувашско-русского перевода // Труды Казанской школы по компьютерной и когнитивной лингвистике. Казань, 2010. – С. 67–71.
6. Желтов П.В. Лингвистические процессоры, формальные модели и методы: Теория и практика. Чебоксары: Изд-во Чуваш. ун-та, 2006.
7. Желтов П.В. Сопоставительно-сравнительное исследование морфем чувашского языка с

применением формальных методов: диссертация ... кандидата филологических наук. Чебоксары, 2010. – 194 с.
8. Исаев Ю.Н. Словообразовательный и семантический анализ флористической терминологии языков различных систем. Чебоксары: Изд-во Чуваш. ун-та, 2010. – 256 с.
9. Ревзин И.И., Юлдашева Г.Д. Грамматика порядков и ее использования // Вопросы языкознания. 1969. № 1. – С. 42–56.

Рецензенты:

Макарычев П.П., д.т.н., профессор, заведующий кафедрой математического обеспечения и применения ЭВМ ФГБОУ ВПО «Пензенский государственный университет», г. Пенза.

Охоткин Г.П., д.т.н., профессор, заведующий кафедрой автоматики и управления в технических системах ФГБОУ ВПО «Чувашский государственный университет имени И.Н. Ульянова», г. Чебоксары.