УДК 81'32

РАЗРЕШЕНИЕ ГРАММАТИЧЕСКОЙ МНОГОЗНАЧНОСТИ В КОРПУСЕ ТАТАРСКОГО ЯЗЫКА

Б.Э. Хакимов, Р.А. Гильмуллин, Р.Р. Гатауллин

Аннотация

В работе рассматриваются вопросы корпусно-ориентированного исследования наиболее частотных типов грамматической омонимии в татарском языке и возможностей автоматизации процесса разрешения многозначности в корпусе. Определяется релевантность генерируемых в процессе автоматического морфологического анализа альтернативных разборов с точки зрения реальной языковой многозначности, предлагаются классификация частотных омоформ и методы разрешения омонимии, оценивается потенциальный эффект для корпуса от разрешения данных типов грамматической многозначности.

Ключевые слова: корпус языка, татарский язык, грамматическая омонимия, омоформа, разрешение многозначности.

Введение

Проблема грамматической многозначности и её разрешения является одной из наиболее актуальных в современной компьютерной и корпусной лингвистике [1]. В национальном корпусе татарского языка «Туган тел» (ТТ), созданном в НИИ «Прикладная семиотика» Академии наук Республики Татарстан совместно с Казанским федеральным университетом, используется система автоматической морфологической разметки на основе собственного морфоанализатора [2]. С целью адекватного отражения специфики татарского языка был разработан морфологический стандарт корпуса [3], ведутся исследования по уточнению и совершенствованию метаязыка описания татарской словоформы [4]. Общая концепция корпуса представлена в [5]. Для разрешения грамматической многозначности в корпусе татарского языка разработчиками предпринято исследование контекстных ограничений различных типов грамматических омонимов с привлечением статистических корпусных данных, предлагаются методы автоматического разрешения грамматической многозначности для татарского языка.

1. Статистические характеристики корпуса татарского языка

На начальном этапе работы из базы текстов корпуса татарского языка (ТТ) были получены статистические данные о частотности словоформ с альтернативными разборами, приведённые в табл. 1. Общий объём корпуса включает 21940452 словоупотребления, доля словоупотреблений с альтернативными разборами составила 25.75%. При этом максимальная длина представленной в корпусе словоформы состоит из основы и двенадцати грамматических аффиксов.

Табл. 1

Словоформы с альтернативными разборами в корпусе татарского языка

Количество альтернативных разборов у вариантных словоформ	Количество словоформ	Доля в корпусе, %
2	4282108	19.52
3	1045392	4.76
4	296547	1.35
5 и более	26773	0.12
Всего словоформ с альтернативными разборами	5650820	25.75

Для выявления наиболее частотных типов омонимии в корпусе и оценки их релевантности с точки зрения реальной языковой омонимии была составлена выборка из 500 наиболее частотных комбинаций альтернативных разборов. На её основе для последующего анализа были отобраны 150 типов, имеющих два варианта разбора, поскольку данная разновидность вариантных словоформ составляет наибольшую долю в корпусе.

2. Оценка релевантности типов омонимов

На первом этапе работы были выделены нерелевантные комбинации омонимов. В данных комбинациях чаще всего альтернативные разборы возникают из-за ошибок морфоанализатора, а именно вследствие избыточности основ в словаре либо в модели словоизменения. Некоторые случаи вызваны некорректными морфонологическими правилами анализатора. Корректировка модели анализатора позволит исключить случаи многозначности, относящиеся к указанным типам.

Отдельный интерес представляют случаи, обусловленные неупотребительностью одного из вариантов разбора. Такие случаи отнесены нами к нерелевантным, так как потенциальные словоформы, автоматически порождаемые в процессе работы морфоанализатора, не представлены в реальном речевом употреблении. Для таких типов необходимо экспериментально определить соответствующее множество словоформ. Всего в корпусной выборке (21940452 словоупотреблений) зафиксировано 5650820 нерелевантных омонимичных разборов (см. табл. 2). Предложенные меры по исключению этих типов омонимии предположительно позволят сократить количество омонимичных разборов на 8.5% (2.1% от общего объёма текстов в корпусе).

3. Классификация частотных типов грамматической омонимии по корпусным данным

Для наиболее частотных лингвистически релевантных типов омонимов была составлена классификация, группирующая отдельные автоматически определённые подтипы (см. табл. 3). Выделены следующие частотные типы омонимов:

1) существительное vs местоимение;
2) глагол vs существительное/прилагательное;
3) местоимение vs числительное;
4) существительное vs прилагательное;

Табл. 2

Нерелевантные типы омонимии[1]

№	1-й омоним	2-й омоним	Частота	Примеры	Причина возникновения	Способ исправления
1	V+PST_DEF	MOD	197779	*иде*	Избыточность в модели анализатора	Исключение одного из вариантов
2	PN	POST	78755	*сыман*, *кебек*, *шикелле*, *сымак*	Избыточность в словаре основ	Исключение варианта PN как некорректного
3	CNJ	INTRJ	56546	*ә*, *йә*	Избыточность в словаре основ	Исключение варианта *х* как некорректного
4	N+NMLZ	N+PSBL	30348	*барлык*, *саклык*, *яктылык*	Избыточность в модели анализатора	Объединение двух вариантов разбора
5	Adj	N+CASE_POINT	23149	*тышкы*, *иске*, *язгы*, *җәйге*, *эчке*, *аскы*	Избыточность в модели анализатора	Исключение варианта N+CASE_POINT
6	Adj+NMLZ	Adj+PSBL	33469	*тереклек*, *авырлык*, *озынлык*, *әзерлек*	Избыточность в модели анализатора	Объединение двух вариантов разбора
7	PN	ADJ	41623	*каршы*, *авыр*	Избыточность в словаре основ	Исключение варианта PN как некорректного
8	N+NMLZ+DIR	N+PSBL+DIR	9417	*барлыкка*, *үсемлеккә*, *казналыкка*, *семьялыкка*, *әсирлеккә*	Избыточность в модели анализатора	Объединение двух вариантов разбора
9	N+ATTR_MUN	N+POSS_1SG+ATTR_MUN	8545	*исемле*, *үтемле*, *керемле*, *дәвамлы*, *төшемле*	Некорректное морфонологическое правило	Исключение варианта N+POSS_1SG+ATTR_MUN
10	V+VN_2+REFL	V+RECP+POSS_3SG+ACC	2787	*торышын*, *табышын*, *үсешен*, *чыгышын*, *таркалышын*, *керешен*	Неупотребительность одного из вариантов разбора	Создание частных правил для определённого множества лексем
	Всего		482418			

[1] Поскольку система грамматической разметки корпуса достаточно сложна, в данной статье приводятся лишь принятые разработчиками корпуса сокращения, а расшифровку и полный список этих сокращений можно найти на сайте корпуса (TT).

5) послелог/местоимение vs существительное/числительное;

6) существительное vs наречие;

7) прилагательное vs существительное с атрибутивным аффиксом;

8) существительное/прилагательное vs существительное с аффиксом принадлежности;

9) прилагательное vs существительное в направительном падеже;

10) прилагательное vs глагол;

11) глагол vs глагол;

12) прилагательное vs наречие;

13) местоимение vs местоимение в местно-временном падеже;

14) существительное vs прилагательное с аффиксом *чА*

15) местоимение vs существительное.

Типы 2, 4, 7, 8, 10–14 представлены множеством регулярно образуемых словоформ, обладающих определённым набором грамматических характеристик. Контекстные правила разрешения многозначности для этих типов обусловлены данными характеристиками и свойствами разрешающего контекста.

Тип 1 представлен лишь одним частотным словом *ул* (‘он/сын’). Тип 3 также представлен лишь одним частотным словом *бер*, употребляемым как в значении числительного ‘один’, так и в функции неопределённого местоимения, близкой к функции неопределённого артикля. Для каждой из частеречных альтернатив действуют свои контекстные закономерности.

Тип 5 включает 4 подтипа, каждый из которых представлен одним словом-послелогом (*өчен* (‘для’), *турында* (‘о’), *буенча* (‘по’)) либо местоимением (*теге* (‘тот’)). Каждое из этих слов имеет омоним-существительное либо числительное в определённой форме. Грамматические характеристики омонимичных слов и синтаксические функции соответствующих послелогов и местоимения определяют контекстные правила для этого типа. Указанные омоформы объединяются в один тип по той причине, что во всех этих случаях второй вариант разбора (существительное/числительное) является крайне маловероятным, а для разрешения многозначности должны использоваться также и статистические методы.

Типы 6 и 9 представлены лексемами *бик* (‘очень/засов’) и *башка* (‘другой/голова+DIR’) соответственно. Тип 15 также является примером омонимии одной словоформы и представлен словом *без* (‘мы/шило’). Данный тип имеет схожие черты с типом 1, однако рассматривается отдельно, так как в случае со словом *без* омонимия распространяется на всю парадигму, а в типе 1 наблюдается лишь в конкретной указанной форме (*ул*).

По всем типам зафиксировано 1636671 омонимичное словоупотребление. Доля в корпусной выборке (21940452 словоупотреблений) составляет 7.4%. Доля среди омонимичных разборов (всего 5650820 по указанной корпусной выборке) – 28.9%.

В данный вариант классификации не включён ещё один особый случай омонимии глагольных форм, связанный с полифункциональностью залоговых аффиксов. Так, статистическое исследование корпусных данных показало, что общее число подобных случаев омонимии в проанализированной выборке текстов составило 408346 словоупотреблений (1.8% от общего объёма текстов и 7.2% от всех альтернативных разборов). Самым частотным подтипом из них является

Табл. 3

Частотные типы грамматической омонимии

№	1-й омоним	2-й омоним	Частота	Примеры
1.	**N**	**PN3_Sing**	**187094**	*ул*
2.	**V**	**N/ADJ**	**591788**	
2.1.	V	N/ADJ	283350	*иң, төр, бет, як, казан, күп, ак, аз*
2.2.	V+FUT_INDF	N+PL	85023	*эшләр, язмалар, кырлар, сулар*
2.3.	V+INF_1	N+PL+DIR	27795	*кырларга, сакларга, чикләргә*
2.4.	V+NEG	N/ADJ	20898	*язма, искәрмә, тукыма, алма*
2.5.	V+PRES	N/ADJ	105946	*буа, тора, әйләнә, тула, кала*
2.6.	V+ADVV_ACC	N	56823	*алып, калып*
2.7.	V+VN_1	N	11953	*басу, сорау, сайлау, чабу*
3.	**PN**	**NUM**	**164687**	*бер*
4.	**N**	**ADJ**	**62935**	*яшь, практик, күк, түгәрәк, чит*
5.	**PP/ Pro**	**NUM/ N**	**163390**	
5.1.	PP	NUM+POSS_3SG +ACC	97023	*өчен*
5.2.	PP	N+POSS_3SG+ LOC	39143	*турында*
5.3.	PP	N+POSS_3SG+ ADV_COMP	15392	*буенча*
5.4.	Pro	N+ POSS_3SG	11832	*теге*
6.	**N**	**ADV**	**85594**	*бик*
7.	**ADJ**	**N+ATTR_MUN**	**42229**	*төрле, тулы, шанлы, исле*
8.	**N/ADJ**	**N+POSS**	**111058**	
8.1.	N	N+POSS_3SG / N+POSS_1SG	86749	*кеше, күбесе, яры, фарсы, теше, җилем, исем, салым, ханым, энем*
8.2.	ADJ	N+POSS_3SG	24309	*туры, каты*
9.	**ADJ**	**N+DIR**	**26082**	*башка*
10.	**ADJ**	**V+ PST_INDF**	**41783**	*узган, кипкән, яткан, үткән*
11.	**V**	**V**	**104983**	
11.1.	V+PST_INDF+ 3PL	V+ADJ_ADVV_ ACC+PL	31162	*тотканнар, дәвалаганнар, охшаганнар, киткәннәр, талаганнар*
11.2.	V	V+PRES	14552	*сына, аша, телә, каба, яна, ура*
11.3.	V+PRES+2SG	V+OBL+ POSS_2SG	24990	*бетерәсең, ышанасың, нишлисең, тотасың, төшерәсең*
11.4.	V+PRES+1SG	V+HOR_SG	34279	*сөйлим, хәтерлим, котлыйм*
12.	**Adj**	**Adv**	**11775**	*якын, хәрәм*
13.	**PN**	**PN+LOC**	**3365**	*биредә, электә*
14.	**N**	**Adj+ADV_ COMP**	**9786**	*яшелчә, сыекча, акча, алача, кызылча*
15.	**PN1_PL**	**N**	**30122**	*без*
Всего			1636671	

V – V+REFL, где одна и та же глагольная словоформа может обозначать как самостоятельную лексему, отдельно включаемую в словарь основ, так и залоговую форму другой лексемы, например: *эзләнергә*, *тотынырга*, *яшеренергә*, *селтәнергә*, *агуланырга*, *алынырга*. Разрешение многозначности данного типа является нетривиальной задачей и во многих случаях требует учёта не только морфо-синтаксических, но и семантических характеристик разрешающего контекста.

4. Автоматическая разработка правил разрешения грамматической многозначности для корпуса татарского языка

Статистические исследования корпуса татарских текстов показывают, что проблема грамматической многозначности для этих текстов стоит довольно остро: четверть (25.75%) словоупотреблений из всего объёма корпуса имеют более одного варианта морфологического разбора (см. табл. 1). Успешное разрешение данного типа многозначности должно в некоторой степени облегчить следующие этапы разрешения многозначности.

При использовании классических методов разрешения грамматической многозначности, основанных на контекстных правилах, необходимо провести классификацию типов омонимов, большую часть которых составляют омоформы. Полная классификация типов омоформ (анализ всего перечня типов) является исключительно трудоёмкой и прагматически неоправданной задачей, так как татарский язык относится к агглютинативным языкам, для которых количество присоединяемых к основе морфем теоретически не ограничено. Например, в указанном корпусе татарских текстов (ТТ) объёмом около 22 млн словоупотреблений число типов омоформ превышает 7000. С другой стороны, использование классических статистических методов осложняется разреженностью данных и отсутствием эталонного размеченного корпуса со снятой омонимией. Таким образом, применение каждого из данных методов не является в достаточной степени эффективным.

Одно из решений этой проблемы описано в [6]. Метод был использован для снятия многозначности в текстах на турецком языке, где количество словоформ, имеющих несколько вариантов разбора, достигает 40%. Согласно результатам данной работы, точность метода для турецкого языка достигла 96% (при точности классических статистических методов в 91%). Типологическая и генетическая близость турецкого и татарского языков даёт основание полагать, что этот метод способен показать хорошие результаты для татарского языка.

Как и в татарском языке, в турецком количество возможных типов многозначности не ограничено, что, в свою очередь, приводит к неудаче при использовании классических статистических методов ввиду разреженности данных. Во избежание этого вместо поиска контекстных ограничений для каждого типа омоформ алгоритм ищет контекстные ограничения для каждой морфемы, количество которых, в отличие от числа типов омоформ, ограничено: для турецкого языка 126 морфем [6], для татарского – 120 морфем [3]. Очевидно, что при таком подходе разреженность данных значительно сокращается.

Согласно описанному методу, для каждой морфемы осуществляется сбор тренировочных данных по выборке словоформ, у которых хотя бы один из

возможных морфологических разборов имеет данную морфему. Полученные данные классифицируются как «положительные» или «отрицательные» в зависимости от того, входит ли морфема в подходящую под контекст парадигму. На основании этих данных с помощью специального алгоритма тренируются правила разрешения грамматической многозначности [6].

Для прогнозирования подходящего варианта разбора незнакомой словоформы морфологический анализатор сначала максимально анализирует словоформы по возможным парадигмам. Далее на основе правил для каждой морфемы определяется некоторая вероятность её присутствия или отсутствия в данной словоформе при данном контексте. Конечный результат рассчитывается с учётом точности каждого правила, и в итоге выбирается самый вероятный разбор [6]. Отличительной особенностью данной модели и используемого алгоритма обучения (GPA algorithm) является их высокая устойчивость к нерелевантным и избыточным признакам.

Проблема отсутствия полностью размеченного корпуса татарского языка со снятой омонимией, который использовался бы в качестве тренировочных данных, может быть частично решена путём выбора для анализа не омоформ с определённой морфемой, а, наоборот, словоформ с данной морфемой и единственным, безальтернативным вариантом разбора. Это позволит выявлять контекстные ограничения непосредственно для морфемы. Однако данный подход покрывает не всё множество морфем (есть, например, морфемы, для которых не обнаружены словоформы с единственно возможным разбором). В подобных случаях контекстные правила разрабатываются вручную либо после полной разметки эталонного фрагмента корпуса.

5. Разработка программных модулей для создания контекстных правил и разрешения грамматической омонимии в корпусе татарского языка

В рамках настоящего исследования разработан программный инструментарий, предназначенный для создания, редактирования и тестирования базы контекстных правил для задач автоматического разрешения грамматической многозначности в татарском языке [7]. Данный модуль можно использовать как отдельно (при этом для всех типов омонимов должны быть построены контекстные правила разрешения), так и в комбинации с вероятностно-статистическими методами. Вторая часть инструментария «LangRuleBase-Модуль РММ» [7] использует эту базу контекстных правил при разрешении грамматической многозначности в текстах. Инструментарий такого рода, принимающий во внимание особенности татарского языка, был разработан впервые и в результате нацелен на облегчение труда исследователя-филолога.

Для облегчения процесса разметки корпуса татарского языка (в том числе и ручного снятия многозначности), а также для обеспечения удобного доступа к статистическим данным корпуса было разработано веб-приложение, обеспечивающее удобство и гибкость работы с корпусом текстов для статистических исследований. Этот программный модуль помимо возможности расширения корпуса и морфологической аннотации поддерживает возможность ручного снятия грамматической многозначности.

Заключение

Формальная контекстно-ориентированная классификация омоформ и разработка контекстных правил разрешения грамматической многозначности с использованием экспериментальных статистических данных в татарском языке осуществляется впервые. Разрабатываемые на основе классификации и контекстных правил лингвистические ресурсы и программные модули позволят разрешать многозначность в корпусе татарского языка и других приложениях. Предполагаемый совокупный эффект в случае разрешения выявленных частотных типов омонимии в корпусе татарского языка может достичь 50%.

Задачами дальнейших исследований являются, с одной стороны, изучение разрешающих контекстов и разработка контекстных правил разрешения многозначности, а с другой – анализ статистических закономерностей в сфере многозначности на разных языковых уровнях и поиск эффективных подходов к разрешению многозначности с учётом специфики татарского языка.

Summary

B.E. Khakimov, R.A.Gilmullin, R.R. Gataullin. Grammatical Disambiguation in the Corpus of the Tatar Language.

This paper deals with the corpus-based study of the most frequent types of grammatical homonymy in the Tatar language and the possibilities to automate disambiguation in the corpus. We determine the relevance of alternative parses generated in the process of automatic morphological analysis in terms of the real language ambiguity. We propose a classification of frequent homoforms and methods for homonymy resolution and also estimate the potential effect of resolution of these types of grammatical ambiguity for the corpus.

Keywords: corpus of a language, Tatar language, grammatical homonymy, homoform, disambiguation.

Источники

TT – Татарский национальный корпус «Туган тел». – URL: http://web-corpora.net/TatarCorpus/search/?interface_language=ru, свободный.

Литература

1. *Невзорова О.А., Зинькина Ю.В., Пяткин Н.В.* Разрешение функциональной омонимии в русском языке на основе контекстных правил // Компьютерная лингвистика и интеллектуальные технологии. – М.: Наука, 2005. – С. 198–202.
2. *Хакимов Б.Э., Гильмуллин Р.А.* К разработке морфологического стандарта для систем автоматической обработки текстов на татарском языке // Системный анализ и семиотическое моделирование. – Казань: Фэн, 2011. – С. 209–214.
3. *Сулейманов Д.Ш., Гильмуллин Р.А.* Двухуровневое описание морфологии татарского языка // Языковая семантика и образ мира: в 2 кн. – Казань: Изд-во Казан. ун-та, 1997. – Кн. 2. – С. 65–67.
4. *Галиева А.М., Хакимов Б.Э., Гатиатуллин А.Р.* Метаязык описания структуры татарской словоформы для корпусной грамматической аннотации // Учён. зап. Казан. ун-та. Сер. Гуманит. науки. – 2013. – Т. 155, кн. 5. – С. 287–296.

5. *Сулейманов Д.Ш., Хакимов Б.Э., Гильмуллин Р.А.* Корпус татарского языка: концептуальные и лингвистические аспекты // Филология и культура. – 2011. – № 4 (26). – С. 211–216.

6. *Yuret D., Türe F.* Learning Morphological Disambiguation Rules for Turkish // Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. – N. Y., 2006. – P. 328–334. – URL: http://dl.acm.org/citation.cfm?id=1220877, свободный.

7. *Сулейманов Д.Ш., Гильмуллин Р.А., Гатауллин Р.Р.* Программный инструментарий для разрешения морфологической многозначности в татарском языке // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS–2014). – Минск: БГУИР, 2014. – С. 503–508.

Поступила в редакцию
25.06.14

Хакимов Булат Эрнстович – кандидат филологических наук, доцент кафедры математической лингвистики и информационных систем в филологии, Казанский (Приволжский) федеральный университет; ведущий научный сотрудник, НИИ «Прикладная семиотика» АН РТ, г. Казань, Россия.

E-mail: *khakeem@yandex.ru*

Гильмуллин Ринат Абрекович – кандидат физико-математических наук, доцент кафедры информационных систем, Казанский (Приволжский) федеральный университет; заведующий отделом когнитивных исследований, НИИ «Прикладная семиотика» АН РТ, г. Казань, Россия.

E-mail: *rinatgilmullin@gmail.com*

Гатауллин Рамиль Раисович – аспирант кафедры информационных систем, Казанский (Приволжский) федеральный университет; младший научный сотрудник, НИИ «Прикладная семиотика» АН РТн, г. Казань, Россия.

E-mail: *ramil.gata@gmail.com*