АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322’366

Ал-др А. Хорошилов, Ю.В. Никитин, С.И. Пшеничный, М.А. Шевкунов, А.А. Хорошилов

Морфологический анализатор МетаФраз нового поколения

Описана концепция, предложены методы и алгоритмы создания морфологического анализатора по технологиям МетаФраз нового поколения, в основу которого, так же как и анализатора МетаФраз первого поколения, положена языковая морфологическая модель флективных классов русских слов, предложенная профессором Г.Г. Белоноговым в конце 60-х гг. прошлого столетия. В концепции содержится ряд положений, обеспечивающих высокое быстродействие и качество обработки текстовых форм слов. Оптимальное соотношение различных типов словарей в составе декларативных средств анализатора, а также применение быстродействующих процедур поиска в этих словарях могут обеспечить требуемые характеристики нового поколения анализатора. В соответствии с предложенными проектными решениями разработан макет морфологического анализатора МетаФраз второго поколения, опытная эксплуатация которого показала его работоспособность и возможность достижения им требуемых технологических и эксплуатационных характеристик.

Ключевые слова: *флективные классы русских слов, морфологический анализатор, машинная грамматика, морфологический анализ, словоизменение, словообразование, программные средства, декларативные средства, автоматическая нормализация слов*

DOI: 10.36535/0548-0027-2021-04-3

ВВЕДЕНИЕ

Стремительный рост объемов текстовой информации в сети Интернет и необходимость обеспечения доступа к ней значительно ускорили развитие технологий машинной обработки текстов на естественном языке (*Natural Language Processing – NLP*) [1]. Эти технологии ориентированы на преобразование текста в его формализованное представление в виде дискретных или комбинаторных структур для дальнейшего выполнения различных аналитических операций. Такое преобразование возможно выполнить с помощью многоступенчатого комплекса процедур морфологического, семантико-синтаксического и концептуального анализа и синтеза текстов. На каждом этапе этого комплекса производится формальное преобразование иерархии единиц смысла текста, выраженных словами, словосочетаниями, предложениями и содержанием текстов. Цель подобного преобразования для каждой единицы смысла – построение формальной модели. Так, например, на этапе обработки слова осуществляется морфологический анализ, обеспечивающий построение его формальной модели на основе информации о буквенном коде. Важной характеристикой процедуры построения морфологической модели является ее быстродействие и точность назначения грамматических характеристик. Эта процедура, как правило, предшествует всем остальными процедурам автоматической обработки текста, и именно она оказывает наибольшее воздействие на процедуры обработки более крупных фрагментов текста.

МАШИННАЯ ГРАММАТИКА НА ОСНОВЕ ФЛЕКТИВНЫХ КЛАССОВ

Под термином «машинная грамматика» понимается комплекс формальных правил, процедур и декларативных средств, обеспечивающих автоматическое преобразование текстовых представлений слов в формальное описание модели в виде совокупности грамматических и семантических характеристик [2]. Обычно машинные грамматики базируются на общепринятых грамматических правилах и формализмах, полученных в результате выявленных закономерностей функционирования языка и речи. Таким формализмом в наших исследованиях является разработанная профессором Г.Г. Белоноговым *система флективных классов слов русского языка* [2, 3]. Её можно рассматривать как универсальную классификацию русских слов, в которой представлены их основные типы словоизменительных парадигм.

Эта классификация положена в основу разработанной нами морфологической модели для русского языка, в рамках которой на основе методов лингвистической аналогии были установлены закономерности между грамматическими характеристиками слов и их конечным буквенным составом, а также предложены методы формального деления слов на слова с регулярной или с аномальной системой словоизменения и словообразования. Такая формальная модель обеспечила возможность автоматического назначения словам грамматических характеристик по их буквенному коду [4].

Система грамматических характеристик классов слов содержится в декларативных средствах, необходимых для функционирования процедур морфологического анализа. Создание декларативных средств больших объемов обеспечивается возможностью как автоматического разделения слов на различные категории, так и автоматического контролируемого назначения им специфических наборов грамматических и семантических характеристик в рамках используемой формальной языковой модели.

В основе концепции и алгоритмических решений морфоанализатора МетаФраз нового (второго) поколения лежит та же языковая модель, что была использована в морфоанализаторе МетаФраз первого поколения [2]. Она базируется на морфологической модели, основанной на системе флективных классов слов и принципе лингвистической аналогии, заложенных в ее реализацию. Принцип лингвистической аналогии опирается на гипотезу, утверждающую, что *в русском языке объективно существует сильная корреляция между конечным буквосочетанием слов и их грамматическими характеристиками.* Для реализации этого принципа разработчиками был создан эталонный словарь конечных буквосочетаний словоформ, представляющий все возможные конечные буквосочетания форм слов в различных контекстных окружениях [5].

При разработке морфологического анализатора МетаФраз первого поколения в качестве основной использовалась приближенная процедура анализа словоформ, выполняемая на основе метода аналогии. При осуществлении этой процедуры было установлено, что назначение грамматических характеристик на основе принципа аналогии возможно только для словоформ, имеющих регулярную трансформационную систему словоизменения. Наличие в словаре конечных буквосочетаний слов с нерегулярной (аномальной) системой словоизменения и словообразования приводило к существенным нарушениям при использовании этого принципа. Поэтому предварительно было необходимо исключить из словаря конечных буквосочетаний слова с аномальной трансформационной системой словоизменения, к которым были отнесены служебные слова, супплетивные формы слов и слова, длина которых не превышала пяти символов. Все они были помещены в словарь, названный словарем коротких и служебных слов.

Ориентация на словари словоформ слов обеспечила высокое быстродействие и точность обработки, поскольку здесь не было необходимости использовать достаточно ресурсоемкие вычислительные действия по установлению основы (или псевдоосновы) и нахождения совместимого грамматического окончания (или псевдоокончания).

Таким образом, в состав словарей морфоанализатора МетаФраз первого поколения первоначально включены только два словаря: словарь коротких и служебных слов и словарь конечных буквосочетаний. В более поздних версиях для большей экономии памяти дополнительно был разработан словарь (таблица) ФКГИ (флективный класс – грамматическая информация), обеспечивающий возможность преобразования базового набора грамматических характеристик[1] в их полный состав и позволяющий существенно сократить объем словарей коротких и служебных слов и конечных буквосочетаний [2].

Обработка слов при помощи алгоритма морфологического анализатора выполняется следующим образом: вначале производится поиск анализируемой формы слова в словаре коротких и служебных слов и, если она там находится, ей назначается базовый набор грамматических характеристик. Если эта словоформа не была обнаружено в словаре коротких и служебных слов, то для нее производится инверсия буквенного состава и выполняется поиск на наибольшее совпадение её конечного буквосочетания с буквосочетанием одного из элементов словаря конечных буквосочетаний. После нахождения подходящего буквосочетания анализируемой словоформе назначаются грамматические характеристики. Недостающая грамматическая информация устанавливается по таблице ФКГИ.

Несколько слов о технологиях создания и ведения этих словарей. Здесь нужно отметить, что относительно простая реализация анализатора, выполненная путем поиска в словаре коротких и служебных слов и словаре конечных буквосочетаний, обеспечивалась сложным комплексом процедур создания и ведения этих словарей. Так, например, если относительно просто выделить короткие и служебные слова, состав которых невелик, то отделить слова, имеющие аномальную систему словоизменительных трансформаций, было значительно сложнее. Это обеспечивалось процедурами анализа их буквенного состава и синтеза всех форм словоизменительных и словообразовательных форм слов, а также процедурами морфемного анализа и проверки совместимости между собой всех типов морфем, составляющих эти словоформы. Достаточно сложной являлась и процедура формирования словаря конечных буквосочетаний, создаваемых на основе больших корпусов текстов. Здесь наряду с задачей определения грамматических признаков было необходимо автоматически определять словоформы с регулярной трансформационной системой словоизменения и словоформы с аномальной трансформационной системой. Для словоформ с регулярной системой словоизменения определялись конечные

[1] К базовым характеристикам словоформы относятся номер флективного класса и грамматическое окончание. Эти характеристики могут обеспечить генерацию более полного набора грамматической информации формы слова: индекс грамматического класса, значений рода, числа, падежа, лица и одушевленности.

буквосочетания, однозначно определяющие грамматические характеристики анализируемой словоформы. Выявленные словоформы с аномальной трансформационной системой переносились в словарь коротких и служебных слов. Для реализации этих задач был разработан комплекс программ в составе Автоматизированной словарной службы.

Необходимо отметить, что концепция МетаФраз, изначально ориентированная на реализацию принципа лингвистической аналогии, была революционной, поскольку предполагалось, что основной поток слов будет обрабатываться не точной процедурой, как это принято в традиционных процедурах морфологического анализа, а приближенной процедурой, которая обычно применялась только для обработки «новых» слов. Точному анализу по словарю словоформ должно подвергаться относительно небольшое число слов, содержащихся в словаре коротких и служебных слов. Еще одним базовым принципом концепции была ориентация на словари словоформ и обработку конкретных форм слов. Был еще и третий принцип – максимальная экономия памяти ЭВМ[2], что объяснялось возможностями ранних этапов развития вычислительной техники. Принципом экономии памяти были пронизаны все применяемые технологические решения, иногда в ущерб качеству и быстродействию. Технология построения словаря конечных буквосочетаний обеспечивала максимальное его сжатие, как по составу элементов, так и по длине эталонных конечных буквосочетаний. Это в значительной степени определяет точность анализа слов на основе принципа лингвистической аналогии. В технологии *создания и ведения словарей* обязательно должна быть операция проверки на совместимость всех элементов морфемной структуры слов в рамках их словоизменительных парадигм.

ДЕКЛАРАТИВНЫЕ СРЕДСТВА МОРФОАНАЛИЗАТОРА МЕТАФРАЗ ВТОРОГО ПОКОЛЕНИЯ

Длительная эксплуатация анализатора МетаФраз первого поколения позволила выявить не только его значительные преимущества, выражающиеся в компактности, быстродействии и высокой точности назначения грамматических характеристик текстовым словам (свыше 95%), но и ряд существенных недостатков. К ним можно отнести относительно небольшой состав грамматических характеристик форм слов, назначаемых в процессе их обработки, а также невозможность автоматизированного назначения словообразовательных и семантических характеристик, присущих всем членам словоизменительной парадигмы слова; например, таких как характеристика глагольности существительных и прилагательных, признак одушевленности существительных, супплетивности прилагательных, а также ряд специфичных признаков принадлежности к группам слов, реализующих важные синтаксические функции при анализе структуры предложения.

Как было отмечено выше, изначальная ориентация на словари словоформ позволила существенно упросить алгоритмы анализа форм слов. Но нужно иметь в виду, что основы слов представляют всю совокупность форм словоизменительной парадигмы, а словари словоформ – только одну форму парадигмы. Поэтому может показаться, что для одинакового покрытия текстов требуется многократное увеличение объемов словарей словоформ. Но, как оказалось, это не совсем так. Наши исследования показали, что при одинаковом покрытии текстов увеличение объемов словарей словоформ может быть не столь значительным. Объемы словарей словоформ и словарей основ слов при одинаковом покрытии текстов находились в отношении 2,5:1, т. е. в среднем в текстах из каждой словоизменительной словоформы содержалось только по 2,5 формы слов из каждой их парадигмы.

Но для существенного повышения покрывающей способности словарей все же словари основ слов предпочтительнее, поскольку они позволяют хранить словообразующую и семантическую информацию для всех членов словоизменительной парадигмы. В связи с этим в концепции МетаФраз второго поколения должна быть предусмотрена возможность использования не только словарей словоформ, но и словарей основ слов без значительного увеличения вычислительной сложности алгоритмов их обработки. Для того, чтобы значительно повысить технологические и эксплуатационные характеристики морфоанализатора МетаФраз нового поколения, необходимо добиться:

1) повышения быстродействия функционирования анализатора не менее чем на 50%;

2) повышения качества декларативных средств анализатора (увеличение совокупности грамматических и семантических характеристик слов до 30 элементов) и значительного увеличения их покрывающей способности;

3) сокращения трудозатрат на разработку декларативных средств и программного обеспечения анализатора не менее чем на 50%.

Возможными путями реализации поставленных задач могут быть следующие решения:

- повышение быстродействия анализатора можно обеспечить включением в его состав дополнительной быстродействующей процедуры, обрабатывающей основной поток (не менее 80%) текстовых словоформ. Такая процедура должна использовать словарь, содержащий полный набор грамматических и семантических характеристик наиболее часто употребляемых форм слов русского языка;
- повышение качества декларативных средств возможно достичь при расширении спектра грамматических (формообразующих и словообразующих) и семантических характеристик. Автоматизированное назначение этих характеристик как словоформам, так и нормальным формам слов должно обеспечиваться контролируемыми технологическими процессами создания и ведения словарей. Формат словарных статей должен быть реализован в виде списковой структуры (название признака – значение признака);

[2] В ранних реализациях объем памяти анализатора не превышал 400 Кбайт.

- повышение покрывающей способности словарей происходит путем включения в состав словарного комплекса словарей словоизменительных парадигм слов, представлять которые будут их нормальные формы;
- сокращение трудозатрат на разработку программных средств и повышение их быстродействия решается за счет разработки относительно простых алгоритмов с небольшой вычислительной сложностью.

Учитывая, что реализация функциональных требований в значительной степени зависит от состава и структуры используемых словарей, дополнительно нами были разработаны частные требования к их составу и структуре, которые обеспечивают:

1) Словарь *S* – назначение полного набора грамматических и семантических характеристик наиболее частотным словоформам;

2) Словарь *K* – назначение усеченного набора грамматических характеристик словоформам с аномальной трансформационной системой словоизменения;

3) Словарь *E* – назначение усеченного набора грамматических характеристик словоформам с регулярной трансформационной системой словоизменения;

4) Словарь *C* – назначение семантических признаков всем членам словоизменительных парадигм слов;

5) Таблица *N* – преобразование текстовой словоформы в ее нормальную форму;

6) Таблица *T* – преобразование усеченного набора грамматических характеристик в их полный набор.

Рассмотрим назначение, лексический состав и набор грамматических и семантических характеристик указанных типов словарей и грамматических таблиц.

Словарь *S* предназначен для обработки основного высокочастотного потока текстовых словоформ. Он включает как все служебные слова, так и наиболее часто встречающиеся формы слов. Обеспечивает назначение набора грамматических и семантических признаков, представляющего более 30 возможных характеристик. Формат словаря – FMA_s30. В табл. 1 приводится фрагмент словаря *S*.

Словарь *K* разработан для обработки остального потока текстовых словоформ. В него включены формы слов, относящиеся к аномальной трансформационной системе словоизменения и словообразования, а также часто встречающиеся формы слов русского языка. Обеспечивает назначение усеченного набора грамматических признаков, состоящего из пяти возможных характеристик. Формат словаря – FMA_k05. В табл. 2 приводится фрагмент словаря *K*.

Словарь *E* используется для обработки словоформ, относящихся к регулярной трансформационной системе словоизменения и словообразования слов русского языка, не покрытых лексикой словарей *S* и *K*. В его состав входят конечные буквосочетания форм слов. Делает возможным назначение усеченного набора грамматических признаков, состоящего из пяти возможных характеристик. .Формат словаря – FMA_k05. В табл. 3 приводится фрагмент словаря *E*.

Словарь *C* создан для обработки представителей словоизменительных парадигм слов и предоставляет возможность назначения всем их членам набора семантических признаков, а также обеспечивает назначение набора словообразовательных и семантических характеристик, состоящего из 18 возможных характеристик. Формат словаря – FMA_c18. В табл. 4 приводится фрагмент словаря *C*.

Таблица *N* служит для преобразования текстовой формы слова в его нормальную форму и содержит нормализующие окончания слов. Включает следующие грамматические признаки: номер флективного класса словоформы и нормализующее окончание, соответствующее номеру флективного класса. Формат таблицы – FMA_n02. В табл. 5 продемонстрирован её фрагмент.

Таблица *T* предназначена для преобразования усеченного набора грамматических признаков словоформ в его полный состав. Обеспечивает назначение полного набора формообразующих характеристик, включающего более восьми возможных характеристик. Формат таблицы – FMA_t08. В табл. 6 приводится её фрагмент.

Таблица 1

Фрагмент словаря *S*

автор	#OK=0#FK=21#GI=*1110#GK=N #OS=QA#TW=1#LI =1#TK=k#RW =w#PD=t#TD=S
администрации	#OK=1#FK=61#GI=*2120*2130*2160*2210*2240#GK=N#SU=t#OS=xf#TK=k#TW=1#TD=S
вашего	#OK=3#FK=114#GI=*1120*1140*3120#GK=m#LI=2#RW=w#OS=QA#TW=2#LI=3#TK=k #PD=t#TK=v #TD=S
кремлевские	#OK=2#FK=106#GI=*0210*0240#GK=A#SU=t#OS=Цg#TW=1#LI=2#TK=k #RW=w#TD=S

Таблица 2

Фрагмент словаря *K*

автоматы	#OK=1#FK=1#TW=1#RW=w#TD=K
автомашину	#OK=1#FK=56#TW=1#RW=w#TD=K
автомеризация	#OK=1#FK=61#TW=1#RW=w #TD=K
автомеханик	#OK=0#FK=31#TW=1#RW=w #TD=K

Таблица 3

Фрагмент словаря *E*

аболз	#OK=1#FK=56#TW=1#RW=w#TD=E
абонзо	#OK=1#FK=1#TW=1#RW=w#TD=E
абор	#OK=1#FK=56#TW=1#RW=w#TD=E
аборок	#OK=1#FK=1#TW=1#RW=w#TD=E

Таблица 4

Фрагмент словаря *C*

автомеханик	#PD=t #TD=C
весь	#DK=e#TD=C
который	#DK=k#TD=C
одна	#DK=0#TD=C
может	#DK=M#TD=C
август	#PT=t#TD=C

Таблица 5

Фрагмент таблицы *N*

#FK=107	#NO=ой
#FK=110	#NO=ой
#FK=111	#NO=ий
#FK=112	#NO=+
#FK=113	#NO=й

Таблица 6

Фрагмент таблицы *T*

#TO=+#FK=011	#GI=*1110*1140#OS=IA#SU=t#GK=N#TD=T
#TO=+#FK=014	#GI=*1110*1140#OS=LA#SU=t#GK=N#TD=T
#TO=+#FK=015	#GI=*1110*1140*1220#OS=MA#SU=t#GK=N#TD=T
#TO=+#FK=016	#GI=*1110*1140#OS=NA#SU=t#GK=N#TD=T
#TO=+#FK=017	#GI=*1110*1140*1220#OS=OA#SU=t#GK=N#TD=T

АЛГОРИТМ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА МЕТАФРАЗ ВТОРОГО ПОКОЛЕНИЯ

Обработка словоформ алгоритмом морфологического анализатора МетаФраз второго поколения выполняется следующим образом: вначале производится прямой поиск анализируемой словоформы по ее буквенному коду в словаре *S* и, если она там находится, ей назначается полный набор грамматических и семантических характеристик и анализ этой словоформы заканчивается. Если эта словоформа не была обнаружена в словаре *S*, то производится прямой поиск по ее буквенному коду в словаре *K*. В случае обнаружения в этом словаре словоформе назначается усеченный набор грамматических характеристик и далее осуществляется преобразование набора усеченной грамматической информации в полный ее состав по таблице *T*. Потом производится назначение анализируемой словоформе семантических признаков по словарю *C*, но предварительно эта словоформа должна быть приведена к ее нормальной форме по таблице *N*.

Словоформы, не обнаруженные в словарях *S* и *K*, обрабатываются по методу лингвистической аналогии на основе анализа их конечных буквосочетаний. Для этого выполняются инверсия буквенного состава словоформы и поиск на наибольшее совпадение ее конечного буквосочетания с буквосочетанием одного из элементов словаря *E*. После нахождения такого буквосочетания анализируемому слову назначается усеченный набор грамматических характеристик. Дальнейшая обработка производится по схеме, аналогичной обработке словоформ, найденных в словаре *K*.

Приведем алгоритм МетаФраз второго поколения.

Алгоритм морфологического анализатора МетаФраз второго поколения

Шаг 1. Выполняется поиск анализируемой формы слова на полное ее совпадение в словаре *S*. В случае успешного поиска словоформе назначается грамматическая и семантическая информация (в соответствии с форматом FMA_s30) и выполняется переход к шагу 7. В случае отсутствия этой словоформы в словаре – переход к шагу 2.

Таблица 7

Результаты работы морфоанализаторана основе технологий МетаФраз второго поколения

00 По	#OK=0#FK=156#GI=*0030#GK=F#OS=ыА#TW=1#TD=S
01 данным	#OK=2#FK=103#GI=*0230*1150*3150#GK=A#OS=ФУ#TW=1#TD=S
02 флота	#OK=1#FK=1#GI=*1120#GK=N#OS=АВ#TW=1#TD=S
03 ,	#OK=0#FK=0#TW=1#GI=*0000#OS=00#GK=,
04 в	#OK=0#FK=164#GI=*0040*0060#GK=F#OS=1А#TW=1#TD=S
05 уходящем	#OK=2#FK=105#TW=1#TD=E#GI=*1160*3160#OS=8S#GK=A#TD=T
06 году	#OK=1#FK=10#GI=*1130*1160#GK=N#OS=Н3#TW=1#TD=S
07 было	#OK=1#FK=125#GI=*3100#GK=L#PG=t#OS=еу#TW=1#TD=S
08 проведено	#OK=1#FK=126#TW=1#TD=K#GI=*3100#OS=еу#GK=K#TD=T#PG=t#TD=C
09 более	#OK=0#FK=152#TW=1#TD=K#GI=*0000#OS=шА#GK=Y#TD=T
10 150	#OK=0#FK=145#GI=*1000#GK=0#OS=уА#TW=1#TD=S
11 учений	#OK=1#FK=73#GI=*3220#GK=N#OS=Йх#TW=1#TD=S
12 различной	#OK=2#FK=103#TW=1#TD=E#GI=*2120*2130*2150*2160#OS=ФВ#GK=A#TD=T
13направленности	#OK=1#FK=55#TW=1#TD=E#GI=*2120*2130*2160*2210*2240#OS=tf#SU=t#GK=N#TD=T
14 ,	#OK=0#FK=0#TW=1#GI=*0000#OS=00#GK=,
15 в	#OK=0#FK=164#GI=*0040*0060#GK=F#OS=1А#TW=1#TD=S
16 ходе	#OK=1#FK=1#GI=*1160#GK=N#OS=АК#TW=1#TD=S
17 которых	#OK=2#FK=103#GI=*0220*0240*0260#GK=k#OS=ФХ#TW=1#TD=S
18 выполнено	#OK=1#FK=126#TW=1#TD=E#GI=*3100#OS=еу#GK=K#TD=T#PG=t#TD=C
19 свыше	#OK=0#FK=155#TW=1#TD=K#GI=*0020#OS=ъА#GK=F#TD=T
20 500	#OK=0#FK=145#GI=*1000#GK=0#OS=уА#TW=1#TD=S
21 боевых	#OK=2#FK=107#GI=*0220*0240*0260#GK=A#OS=ЧХ#TW=1#TD=S
22 упражнений	#OK=1#FK=73#TW=1#TD=E#GI=*3220#OS=Йх#GK=N#TD=T
23 и	#OK=0#FK=153#GI=*0000#GK=&#OS=ЩА#TW=1#TD=S
24 применений	#OK=1#FK=73#TW=1#TD=E#GI=*3220#OS=Йх#GK=N#TD=T
25 оружия	#OK=1#FK=73#GI=*3120*3210*3240#GK=N#SU=t#OS=Йм#TW=1#TD=S
26 .	#OK=0#FK=0#TW=1#GI=*0000#OS=00#GK=.

Шаг 2. Выполняется поиск анализируемой словоформы на полное ее совпадение в словаре *K*. В случае успешного поиска ей назначается усеченная грамматическая информация (в соответствии с форматом FMA_k5) и выполняется переход к шагу 4. В случае отсутствия этой словоформы в словаре – переход к шагу 3.

Шаг 3. Производится инверсия буквенного состава анализируемого слова и выполняется поиск конечного буквосочетания анализируемой формы слова на наибольшее совпадение с одним из элементов словаря *E*. В случае успешного поиска словоформе назначается усеченная грамматическая информация (в соответствии с форматом FMA_k5) и выполняется переход к шагу 4.

Шаг 4. Выполняется поиск в таблице *T* по двум грамматическим характеристикам – номеру флективного класса и текстового грамматического окончания. В случае успешного поиска словоформе назначается полный набор грамматической информации (в соответствии с форматом FMA_t8) и выполняется переход к шагу 5.

Шаг 5. По таблице *N* выполняется приведение текстовой словоформы к ее нормальной форме путем присоединения к словоизменительной основе нормализующего окончания, соответствующего номеру флективного класса(в соответствии с форматом FMA_n2). По завершению операции выполняется переход к шагу 6.

Шаг 6. Выполняется прямой поиск сформированной на шаге 5 нормальной формы слова на полное его совпадение в словаре *C*. В случае успешного поиска всем членам словоизменительной парадигмы назначается семантическая информация (в соответствии с форматом FMA_c18) и выполняется переход к шагу 7. В случае отсутствия представителя словоизменительной парадигмы в словаре – переход к шагу 7.

Шаг 7. Выполняется преобразование полученных результатов в структуру метаданных.

В табл. 7 приводятся результаты работы анализатора на основе технологий МетаФраз второго поколения.

ХАРАКТЕРИСТИКИ ПРОГРАММНЫХ И ДЕКЛАРАТИВНЫХ СРЕДСТВ

При проектировании анализатора на основе технологий МетаФраз второго поколения должны быть предварительно установлены параметры вычислительной сложности алгоритмов[3] отдельных процедур с целью определения их быстродействия при реализации программного кода. Для упрощения этой зада-

[3]Под термином «вычислительная сложность алгоритма» в информатике и теории алгоритмов понимается функция зависимости объема работы, которая выполняется некоторым алгоритмом, от размера входных данных. Объем работы обычно измеряется абстрактными понятиями времени и пространства, называемыми вычислительными ресурсами.

чи за единицу вычислительной сложности процедуры примем прямой поиск в хешированном массиве, при этом сопутствующие вычислительные действия, связанные с преобразованием данных, также включим в эту единицу сложности вычислений. Операцию обратного поиска на наибольшее вхождение оценим в пять единиц.

Суммарные результаты вычислительной сложности каждой технологической цепочки операций морфоанализатора МетаФраз первого поколения приведены в табл. 8, а в табл. 9 – аналогичные результаты вычислительной сложности каждой технологической цепочки операций морфоанализатора на основе технологий МетаФраз второго поколения.

Анализ табл. 8 и 9 показывает, что минимальная вычислительная сложность технологической цепочки №1 МетаФраз первого поколения равна двум, а цепочки №2 – семи. Для анализатора на основе технологий МетаФраз второго поколения этот диапазон значений более широкий. Так, вычислительная сложность технологической цепочки №1 анализатора здесь равна единице, вычислительная сложность технологической цепочки №2 – пяти, а вычислительная сложность технологической цепочки №3 – 10.

Таблица 8

Суммарные результаты вычислительной сложности каждой технологической цепочки операций в морфоанализаторе МетаФраз первого поколения*

Технологич. цепочки операций	Поиск в словарях и грамматических таблицах			Суммарная вычислительная сложность технологических цепочек операций
	КСС	КБС	ФКГИ	
№1	1	–	1	2
№2	1	5	1	7

* 1. Технологической операцией цепочки №1 является а) прямой поиск в словаре коротких и служебных слов (КСС) и б) поиск в словаре флективный класс – грамматическая информация (ФКГИ).

2. Технологической операцией цепочки №2 является а) прямой поиск в словаре коротких и служебных слов (КСС), б) обратный поиск в словаре конечных буквосочетаний и в) поиск в словаре флективный класс – грамматическая информация (ФКГИ).

Таблица 9

Суммарные результаты вычислительной сложности каждой технологической цепочки операции в морфоанализаторе МетаФраз второго поколения*

Технологич. цепочки операций	Поиск в словарях и грамматических таблицах						Суммарная вычислительная сложность технологических цепочек операций
	S	*K*	*E*	*T*	*N*	*C*	
№1	1	–	–	–	–	–	1
№2	1	1	–	1	1	1	5
№3	1	1	5	1	1	1	10

***** 1. Технологической операцией цепочки №1 является а) прямой поиск в словаре S.

2. Технологической операцией цепочки №2 является а) прямой поиск в словаре S, б) прямой поиск в словаре K, в) прямой поиск в таблице T, д) прямой поиск в таблице N, г) прямой поиск в словаре C

3. Технологической операцией цепочки №3 является а) прямой поиск в словаре S, б) прямой поиск в словаре K, в) обратный поиск в словаре E, д) прямой поиск в таблице T, г) прямой поиск в таблице N, е) прямой поиск в словаре C.

Таблица 10

Параметры комплекса словарей МетаФраз первого поколения

Тип словаря	Объем словаря	Эффективный объем словоформ	Количество граммат. признаков	Вероятность правильного назначения грамматической информации	Генерация словоформ в текстах сверхбольшого объема	Покрытие текстов каждым словарем, %
Словарь коротких и служебных слов	78000	78000	4	100%	13731372	53
Словарь конечных буквосочетаний	44000	3–7 млн	4	77%	1–3 млн	100

Таблица 11

Параметры комплекса словарей МетаФраз второго поколения

Тип словаря	Объем словаря	Эффективный объем словоформ	Количество грамматич. и семантич. признаков	Вероятность правильного назначения грамматической информации, %	Встречаемость словоформ в текстах большого объема	Покрытие текстов каждым словарем, %
Словарь S	40000	40000	34	100	26078948	81
СловарьK	120000	120000	8	100	27431372	97
Словарь E	150000	3–20 млн	8	87	3–20 млн	100
Словарь C	120000	600000	26	100	27414081	99

Очевидно, что вычислительная сложность алгоритмов и процедур МетаФразв значительной степени зависит от соотношения объемов текстовых слов, которые будут обрабатываться каждой технологической цепочкой операций. Поэтому увеличение потока текстовых слов, обрабатываемых технологической цепочкой №1, обеспечит снижение общей сложности обработки процедур МетаФраз и, соответственно, приведет к увеличению ее быстродействия. Таким образом, видно, что количественные параметры комплекса словарей МетаФраз напрямую влияют на их быстродействие.

Ниже приведены количественные параметры комплекса словарей МетаФраз первого (табл. 10) и второго (табл. 11) поколений, иллюстрирующие динамику изменения и перераспределения количественного состава словарей МетаФраз.

ТЕХНОЛОГИИ СОЗДАНИЯ И ВЕДЕНИЯ ДЕКЛАРАТИВНЫХ СРЕДСТВ

Очевидно, что создание необходимых объемов разного типа словарей и большое количество сопутствующих грамматических и семантических характеристик их элементов невозможно реализовать ручными методами. Для этого требуются программно-лингвистические средства автоматизации создания словарей и грамматических таблиц, которые в рамках концепции фразеологического анализа текстов принято называть Автоматизированной словарной службой[4] (АСС). Кратко определим объекты и средства автоматизации в рамках АСС.

Под *объектами автоматизации* АСС будем понимать массивы групп и подгрупп словоформ русского языка, расклассифицированные по совокупности грамматических и семантических характеристик. Общее число таких групп и подгрупп словоформ будет составлять несколько десятков.

Средствами автоматизации АСС обозначим технологические операции, приводящие к трансформации их буквенного кода и состава грамматических и семантических признаков.

Каждое текущее значение состава грамматических и семантических признаков назовем *форматом словоформы*.

К *технологической операции* АСС отнесем локальное изменение формата словоформы из одного состояния в другое.

Таким образом, основная задача реализации *технологий АСС* – это автоматизированное выполнение цепочки технологических операций с целью *создания комплекса словарей для морфологического анализа МетаФраз второго поколения*, при минимальном участие человека в этом процессе.

Источниками информации для формирования декларативных средств могут служить следующие лингвистические ресурсы.

1. Для формирования словаря *S* источником являются частотные словари словоформ, созданные на больших корпусах текстов. Необходимо обработать и включить в состав словаря *S* частотную часть этих словарей, статистическая информация об одном из которых содержится в табл. 12.

2. Для формирования словаря *K* источником служит словарь коротких и служебных слов МетаФраз первого поколения. Необходимо будет выполнить переформатирование грамматических характеристик словаря.

3. Источник информации для формирования словаря *E* – словарь конечных буквосочетаний МетаФраз первого поколения. Необходимо будет выполнить переформатирование грамматических характеристик словаря.

4. Словарь *C* формируется на основе частотных словарей представителей словоизменительных парадигм, созданных на больших корпусах текстов. Необходимо обработать и включить в состав словаря *C* частотную часть этих словарей, статистическая информация об одном из которых содержится в табл. 13.

Основные технологические операции Автоматизированной словарной службы

1. Назначение базовых грамматических характеристик словоформам на основе методов лингвистической аналогии.

2. Автоматическое назначение полного набора грамматических характеристик на основе анализа их базовых характеристик.

[4]Автоматизированная словарная служба (АСС) – это сложный программно-информационный комплекс, обеспечивающий возможность автоматизированной интеллектуальной обработки текстовой информации с целью ее преобразования в систему словарных конструкций, сопровождаемых совокупностью их грамматических и семантических характеристик.

3. Вычисление совокупности одних семантических характеристик на основе анализа других характеристик.

4. Шаблонное назначение семантических характеристик, присущих конкретной группе слов.

Как отмечалось, основным источником пополнения словарей *S* и *C* была лексика тематических корпусов текстов, которая выбиралась по статистическим параметрам. Для получения этих параметров были сформированы частотные словари двух типов: словарь словоформ и словарь словоизменительных парадигм. В частотном словаре словоизменительных парадигм каждая парадигма представлена нормализованной формой слова парадигмы.

Таблица 12

Статистические данные о частотном словаре словоформ, составленном по корпусу текстов общим объемом 28,5 млн слов

Ранги частот	Макс.частота диапазона словоформ	Мин.частота диапазона словоформ	Количество разных словоформ в корпусе текстов	Общее количество словоформ в корпусе текстов	Покрытие корпуса текстов словоформами
1	1163633	22988	100	9033294	0,316575
2	22988	5696	500	12977269	0,454793
3	5696	3140	1000	15064240	0,527932
4	3140	1645	2000	17286279	0,605804
5	1645	1110	3000	18625593	0,652740
6	1110	833	4000	19583726	0,686319
7	833	633	5000	20329325	0,712448
8	633	303	10000	22547206	0,790175
9	303	185	15000	23731372	0,831674
10	185	127	20000	26422951	0,858814
11	127	73	30000	25474618	0,892767
12	73	48	40000	26078948	0,913946
13	48	11	50000	27161190	0,927571
14	11	11	100000	27500566	0,963767
15	11	4	194461	28046791	0,982910
16	4	2	305591	27552620	0,992041
17	2	1	532706	28534454	1,000000

Количество разных словоформ в словаре равно 532 706.

Таблица 13

Статистические данные о частотном словаре нормализованных форм слов словоизменительных парадигм, составленном по корпусу текстов общим объемом 28,5 млн слов

Ранги частот	Макс. частота диапазона словоизм. парадигм	Мин. частота диапазона словоизм. парадигм	Количество разных словоизм. парадигм в корпусе текстов	Обще количество словоизм. парадигм в корпусе текстов	Покрытие корпуса текстов словоизм. парадигм
1	1164325	30960	100	10560385	0,383281
2	30960	7113	500	15848771	0,575218
3	7113	3743	1000	18385769	0,667297
4	3743	1752	2000	20900122	0,758553
5	1752	1068	3000	22254899	0,807724
6	1068	726	4000	23132119	0,839562
7	726	532	5000	23753139	0,862101
8	532	191	10000	25340828	0,919725
9	191	99	15000	26032609	0,944832
10	99	60	20000	26422951	0,959000
11	60	28	30000	26840747	0,974163
12	28	16	40000	27053112	0,981871
13	16	11	50000	27161190	0,985793
14	11	3	100000	27414081	0,994972
15	3	2	125444	27468955	0,996963
16	2	1	209109	27552620	1,000000

Количество разных словоизменительных парадигм слов в словаре равно 209 109.

Словарь первого типа служил источником для пополнения словаря *S*. Словарь второго типа – источником для пополнения словаря *C*. Статистические данные о частотном словаре форм слов корпуса текстов, общим объемом 28 534 454 слов, приведены в табл. 12. Статистические данные о частотном словаре представителей словоизменительных парадигм этого корпуса текстов приведены в табл. 13.

На основе имеющихся статистических данных о конкретных словоформах (табл. 12) и статистических данных о нормализованных формах слов, представляющих словоизменительные парадигмы (табл. 13), возможно принимать решения по формированию и пополнению словарей *S* и *C* частотной лексикой.

ЗАКЛЮЧЕНИЕ

По итогам создания морфологического анализатора на основе технологий МетаФраз второго поколения можно отметить следующее.

1. В основу анализатора была положена морфологическая модель, основанная на системе флективных классов русского языка.

2. Система флективных классов, разработанная профессором Г.Г. Белоноговым в конце 60-х гг. прошлого века, представляет многообразие типов словоизменения русского языка.

3. Принцип лингвистической аналогии дает возможность обеспечить назначение грамматических характеристик словоформ бессловарным методом, основанным на анализе конечных буквосочетаний слов, а также позволяет сократить трудозатраты при формировании декларативных средств.

4. Повышение быстродействия предлагаемого морфоанализатора за счет включения в его состав дополнительной быстродействующей процедуры, обрабатывающей основной поток (не менее 80%) текстовых словоформ.

5. Повышение качества декларативных средств путем расширения спектра грамматических (формообразующих и словообразующих) и семантических характеристик.

6. Автоматизация назначения грамматических и семантических характеристик как словоформам, так и нормальным формам слов.

7. Увеличение покрывающей способности словарей при включении в состав словарного комплекса наряду со словарями словоформ словарей словоизменительных парадигм слов в виде их нормальных форм.

8. Сокращение трудозатрат на разработку программных средств и повышение их быстродействия путем разработки относительно простых алгоритмов с небольшой вычислительной сложностью.

В заключение определим назначение и местоположение в составе базовых средств семантико-синтаксического анализа текстов процедуры разрешения омонимии[5]. На наш взгляд разрешение как лексической (частеречевой), так и семантической омонимии невозможно выполнить без учета контекста. Между тем процедура морфологического анализа ориентирована на анализ слов вне контекста. Поэтому в предлагаемом морфоанализаторе МетаФраз второго поколения эта операция не предусмотрена. Ее реализация возможна только на этапе семантико-синтаксического или концептуального анализа [2, 6], когда появляется возможность опираться на контекст омонимичной формы слова, но при этом предварительно морфологический анализатор должен предоставить информацию о наличии омонимии у анализируемой формы слова.

СПИСОК ЛИТЕРАТУРЫ

1. Белоногов Г.Г., Гиляревский Р.С., Селедков С.Н., Хорошилов Ал-р А. О путях повышения качества поиска текстовой информации в системе Интернет // Научно-техническая информация. Сер. 2. – 2013. – № 8. – С. 15–22; Belonogov G.G., Gilyaresvkii R.S., Seletkov S.N., Khoroshilov A.A. Ways to Improve the Quality of Textual Data Searches on the Internet. – Automatic Documentation and Mathematical Linguistics. – 2013. – Vol. 47, № 4. – P. 111-120.
2. Аблов И.В., Козичев В.Н., Ширманов А.В., Хорошилов Ал-р А., Хорошилов Ал-ей А. Средства машинной грамматики русского языка (по Г.Г. Белоногову) // Научно-техническая информация. Сер. 2. – 2018. – № 6. – С. 32-46.
3. Белоногов Г.Г., Калинин Ю.П., Хорошилов Ал-др А., Хорошилов Ал-ей А. Компьютерная лингвистика и перспективные информационные технологии // Научно-техническая информация. Сер. 2. – 2004. – № 8. – С. 22–32.
4. Старовойтов А.В., Пошатаев О.Н., Прохоров С.Н., Хорошилов Ал-р А. Методы автоматизированного составления и ведения словарей // Информатизация и связь. – 2013. – №3. – С. 91–97.
5. Белоногов Г.Г., Зеленков Ю.Г., Новоселов А.П., Хорошилов Ал-др А., Хорошилов Ал-ей А. Метод аналогии в компьютерной лингвистике // Научно-техническая информация. Сер. 2. – 2000. – № 1. – С. 21–31.
6. Кан А.В., Ревина В.Д., Руснак В.И., Хорошилов Ал-др А., Хорошилов Ал-сей А. Автоматическоеформирование синтаксической модели языка для задач машинного перевода и информационного поиска // Научно-техническая информация. Сер. 2. – 2018. – № 12. – С. 25-41.

Материал поступил в редакцию 08.02.21.

[5] Академик В.В. Виноградов в статье «Об омонимии и смежных с ней явлениях» (журнал «Вопросы языкознания» 1968 г.) определил это лингвистическое понятие как «..звуковое и грамматическое совпадение языковых единиц, которые семантически не связаны друг с другом..»

Сведения об авторах

ХОРОШИЛОВ Александр Алексеевич – доктор технических наук, профессор НИУ МАИ; ведущий научный сотрудник Федерального исследовательского центра Информатики и Управления (ФИЦ ИУ) РАН; старший научный сотрудник 27 ЦНИИ Министерства обороны РФ, Москва
e-mail: khoroshilov@mail.ru

НИКИТИН Юрий Викторович – научный сотрудник ФИЦ ИУ РАН, руководитель группы разработки АО «НПК «ВТ и СС»
e-mail: yuri.v.nikitin@gmail.com

ПШЕНИЧНЫЙ Сергей Игоревич – кандидат экономических наук, директор программ АО «НПК «ВТ и СС»
e-mail: s.pshenichniy@htsts.ru

ШЕВКУНОВ Максим Александрович – ведущий конструктор АО «НПК «ВТ и СС»
e-mail: mshevkunov@htsts.ru

ХОРОШИЛОВ Алексей Алексеевич – кандидат технических наук, старший научный сотрудник 27 ЦНИИ Министерства обороны РФ, Москва
e-mail: alex_khoroshilov@mail.ru