УДК 004.934.5

Л.И. Цирульник[1], В.В. Веремей[2]

АЛГОРИТМЫ СОЗДАНИЯ И ПОПОЛНЕНИЯ ГРАММАТИЧЕСКОГО СЛОВАРЯ РУССКОГО ЯЗЫКА ДЛЯ СИНТЕЗА РЕЧИ ПО ТЕКСТУ

Описываются процедуры создания и пополнения грамматического словаря. Показываются особенности словесных ударений и дополнительные грамматические характеристики, вносимые в словарь и используемые для синтеза речи по тексту. Проводится эксперимент по оценке полноты покрытия словарем текстов различных жанров. Предлагается алгоритм автоматизированного пополнения словаря.

Введение

Система синтеза речи по тексту состоит из двух последовательных этапов: обработки входного текста и обработки речевого сигнала [1–4]. Обработка текста включает, в свою очередь, блок интонационной разметки и блок фонетического преобразования. В блоке интонационной разметки осуществляется членение предложений на фразы, членение фраз на просодические синтагмы, определение словесного ударения, деление синтагм на акцентные единицы и определение интонационного типа синтагм. Очевидно, что для корректной интонационной разметки необходим анализ входного текста, который, в принципе, может быть более или менее глубоким – от лексико-грамматического анализа до семантического или прагматического. При этом чем глубже уровень анализа текста, тем более выразительную и естественную интонацию можно синтезировать. Для осуществления любого из перечисленных видов анализа необходима информация о лексико-грамматических категориях слов входного текста, не менее важна при интонационной разметке и информация о позиции ударения слов входного текста. И та и другая информация содержится в грамматическом словаре, который, таким образом, должен присутствовать и использоваться в системе синтеза речи по тексту.

К настоящему времени создан ряд электронных грамматических словарей, наиболее полный из них [5] содержит более 97 000 парадигм слов. Все созданные словари основаны на «Грамматическом словаре русского языка» А.А. Зализняка [6], иногда с некоторыми дополнениями, связанными со спецификой решаемых задач обработки текста. Однако существующие словари не могут использоваться в системе синтеза речи по тексту, поскольку, во-первых, большинство из них не содержит информации о позиции ударения в словах (для многих задач обработки текста эта информация не требуется) и, во-вторых, ни один из них не содержит дополнительной информации, необходимой при обработке текста для синтеза речи, например о наличии слабых ударений в некоторых служебных частях речи, наличии расчленяемых частиц и т. п. Кроме того, любой существующий словарь является неполным и требует определенного пополнения. Особенности создания и пополнения грамматического словаря для синтеза речи по тексту показаны в данной статье.

1. Процедура создания словаря

Для разрабатываемой системы синтеза русской речи необходима база всех слов русского языка с указанными в них позициями ударения и набором лексико-грамматических категорий (ЛГК), относящихся к каждой словоформе. В качестве источника этих сведений был использован «Грамматический словарь русского языка» А.А. Зализняка, в котором информация о грамматической парадигме слова (под парадигмой понимается совокупность всех грамматических форм некоторого слова) дается с помощью системы условных обозначений и индексов, описывая правила словоизменения. В соответствии с этими правилами и были сгенерированы словоформы каждой парадигмы словаря.

1.1. Основные характеристики грамматического словаря А.А. Зализняка

Грамматический словарь представлен в виде текстового документа, содержащего словарные статьи. Каждая словарная статья состоит из заглавного слова (исходная форма), грамматических категорий, индекса, а также дополнительных сведений (в частных случаях один из этих элементов может отсутствовать).

Все грамматические категории можно разделить на следующие виды:

– несловоизменительные, значения которых одинаковы для всех словоформ парадигмы (например, категория рода существительных и категория вида глагола);

– словоизменительные, изменение значений которых приводит к образованию новой словоформы в парадигме (например, категория падежа и числа существительного).

Служебные части речи (предлог, союз, частица, междометие и т. п.) обладают категориями только первого вида, в то время как категории большинства знаменательных частей речи (имя существительное, имя прилагательное, глагол и т. п.) представлены обоими видами (рис. 1). В словарной статье содержатся несловоизменительные категории, также указывающие на часть речи, которой является рассматриваемое слово.

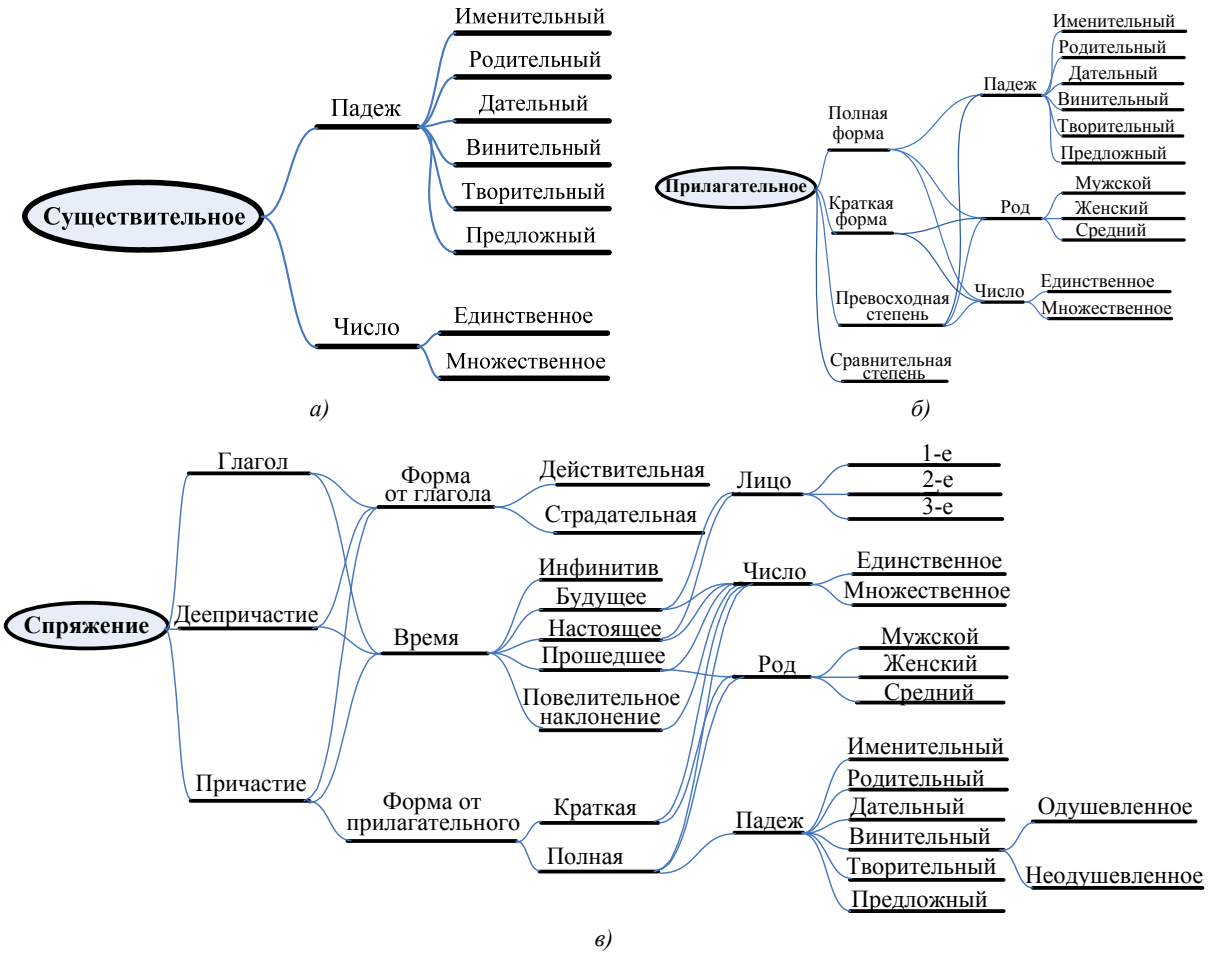

Рис. 1. Словоизменительные категории: *а)* существительного; *б)* прилагательного; *в)* глагола

Индекс имеется только у изменяемых частей речи. Он определяет правила построения парадигмы, а также схему ударения в ней.

В качестве дополнительных сведений указываются некоторые чередования при образовании словоформ или некоторые часто встречающиеся отклонения от стандартных склонений или спряжений. Иногда указываются явно некоторые формы, образующиеся не по правилам.

Рассмотрим примеры построения парадигм слов, представленных в словарных статьях <конкурс м 1а> и <благоволить нсв нп 4в>. Здесь в соответствии с описанием словарной

статьи, данным выше, «конкурс» – это исходная форма, «м» – грамматическая категория, «1а» – индекс; дополнительные сведения в данной статье отсутствуют. Аналогично в статье <благоволить нсв нп 4в> «благоволить» – исходная форма, «нсв нп» – грамматическая категория, «4в» – индекс. Символ “м” (в словарной статье «конкурс») указывает на то, что слово является существительным, неодушевленным, мужского рода, субстантивного склонения. Совокупность обозначений “нсв нп” (в словарной статье «благоволить») указывает, что слово является непереходным глаголом несовершенного вида. Символ “1” в контексте существительного обозначает тип склонения, зависящий от окончания основы слова. Символ “4” в контексте глагола обозначает тип спряжения, зависящий от финальной части основы и характеризуемый финалиями инфинитива *-ить* и окончаниями форм глагола 1-го, 2-го и 3-го лица единственного числа настоящего времени соответственно: *-ю (-у)*, *-ишь*, *-ит*. Символы “а” и “в”, входящие в состав индекса, обозначают схему ударения в парадигме: постоянное ударение на основе слова и постоянное ударение на окончании слова соответственно.

1.2. Особенности создания грамматического словаря для системы синтеза речи по тексту

В грамматическом словаре слова в основном делятся на грамматические разряды. Под грамматическим разрядом понимается совокупность слов, у которых набор ЛГК одинаков. Например, слова *рука*, *слон*, *сердце*, *запятая* изменяются по падежам и числам, следовательно, они относятся к одному и тому же грамматическому разряду. Таким образом, все существительные составляют один грамматический разряд.

Прилагательные делятся на два грамматических разряда: изменяются по падежам, числам, родам, категории одушевленности–неодушевленности – один грамматический разряд и имеют, в дополнение к первому разряду, краткие формы – второй грамматический разряд. Сведения о правилах построения сравнительных степеней даются в словарной статье, превосходные же степени сравнения представлены как самостоятельные прилагательные, однако в словарь они включены выборочно. По этой причине было принято решение не обрабатывать словарные статьи, содержащие превосходные степени, а генерировать эти словоформы при построении парадигмы соответствующего прилагательного начальной формы (единственного числа, мужского рода, именительного падежа), учитывая, что существование превосходной степени в парадигме определяется, как правило, существованием сравнительной степени. При такой генерации использовалось следующее правило: образование формы превосходной степени происходит путем добавления к основе прилагательного суффикса *-айш,* если она оканчивается на *г*, *к*, *х* (при этом учитывается соответствующее чередование согласных) и *-ейш* во всех остальных случаях (например, *легко – легчайший*, *быстрый – быстрейший*).

Числительные в словарных статьях грамматического словаря содержат одну ЛГК (часть речи) и множество дополнительных помет, в которых отражены правила склонения. Таким образом, общие правила склонения числительных в грамматическом словаре не представлены, что сделало практически невозможной автоматизацию генерации словоформ числительных. По этой причине все числительные, их склонения, а также дополнительные лексико-грамматические характеристики были добавлены в словарь вручную. При добавлении обозначалась такая ЛГК, как тип числительного, по которой числительные бывают количественные (*один*, *полтора*), порядковые (*восьмой*, *сотый*) и собирательные (*двое*, *десятеро*). Кроме того, были введены характеристики рода (*полтора* раза – *полторы* страницы, *сотый* поступок – *сотая* ошибка), числа (*один – одни, первый – первые*), а также одушевленности (увидеть *второго* человека – отремонтировать *второй* стол).

В грамматическом словаре местоимения представлены двумя грамматическими разрядами: местоимение-существительное (*я*, *мы*, *кто*, *кое-что*) и местоимение-прилагательное (*наш*, *этот*, *некоторый*), которые показывают тип словоизменения местоимения. В задачах синтеза речи такая классификация недостаточна. По этой причине словарь был пополнен местоимениями, которые дополнились лексико-грамматическими категориями. Так был выделен семантический тип, по которому местоимения делятся на возвратные (*себя*), возвратно-притяжательные (*свое*, *своему*), вопросительно-относительные (*какой*, *чей*, *который*), личные (*я*, *вы*), неопределенные (*некий*, *кто-либо*), определительные (*самый*, *все*, *каждый*), отрицательные (*ничто*, *никакой*, *ничей*),

притяжательные (*мой*, *его*), указательные (*тот*, *таковой*). Как правило, каждое местоимение также имеет следующие ЛГК: число, род и падеж, а личные и притяжательные местоимения можно классифицировать и по лицам.

В глагольную парадигму включаются как личные, так и неличные формы (причастия и деепричастия). По причине большого количества причастных форм причастия и деепричастия выделены в отдельную таблицу в базе словаря и отнесены к самостоятельным частям речи.

Все прочие слова, неизменяемые части речи (предлоги, союзы, частицы, вводные слова, предикативы (слова в значении сказуемого, например *ветрено*), междометия и наречия) представляют собой один грамматический разряд в соответствии со словоизменением. Однако для задач синтеза речи таких сведений для некоторых неизменяемых частей речи недостаточно. Например, для синтеза речи важно (в грамматическом словаре не представлено), что союзы бывают не только простые, но и составные, состоящие из двух и более слов (например, *между тем как*, *то… то*, *не то… не то*). Составные союзы делятся, в свою очередь, на повторяющиеся (*то ли … то ли*), расчлененные (*не только … но и*) и, собственно, составные (*а именно*, *то есть*). Словарь был пополнен составными союзами, а также их характеристиками. При обработке текста в процессе синтеза речи имеет большое значение определение союзов (особенно составных) для правильного указания позиций ударения в них, а также для установки определенного типа интонации.

Для синтеза речи важно деление частиц на нерасчленяемые (*так-таки*, *вряд ли*), расчленяемые (*хоть бы*, *едва ли не*) и составные (*что из того что*).

Большое значение при синтезе речи имеет определение предлога как первообразного или производного. Первообразный предлог может сочетаться с тремя падежами (*по*, *с*), с двумя падежами (*в*, *за*, *между*) или с одним падежом (*без*, *для*, *до*). Производные предлоги – это предлоги, образованные от знаменательных слов – существительных, наречий или глаголов (деепричастий). Таким образом, производные предлоги делятся на отыменные (*ввиду*, *в качестве*, *во имя*), наречные (*близ*, *сверх*, *после*) и отглагольные (*включая*, *исключая*, *не считая*). Каждый из производных предлогов соединяется только с одним каким-нибудь падежом. Как и союзы, предлоги могут быть простыми (*из-под*, *изнутри*) и составными (*в качестве*, *вместе с*).

Для синтеза речи также важны лексико-грамматические категории наречий, которые не представлены в грамматическом словаре. Все наречия делятся на местоименные (*вовсю*, *где-либо*) и знаменательные (*волчком*, *ва-банк*). Местоименные наречия делятся, в свою очередь, на личные (*по-твоему*, *по-нашему*), возвратные (*по-свойски*), указательные (*здесь*, *там*, *туда*), определительные (*всячески*, *везде*, *много*), вопросительные (*где*, *куда*, *зачем*), неопределенные (*где-то*, *куда-либо*), отрицательные (*нигде*, *никуда*, *никогда*). По лексическому значению все наречия, как знаменательные, так и местоименные, делятся на определительные (*волей-неволей*, *бегом*) и обстоятельные (*около*, *поверху*).

Большое значение для синтеза речи имеет маркировка ударений в словах. При этом различаются три типа слов:

полноударные – слова, имеющие одно полное, или сильное, ударение;

частично ударные – слова, имеющие частичное, или слабое, ударение. В процессе синтеза речи такие слова присоединяются к предыдущему либо последующему, образуя одну акцентную единицу (группу слов с одним сильным ударением);

клитики – слова, которые в слитной речи произносятся без ударения и присоединяются к предыдущему (энклитики) либо последующему (проклитики) словам, образуя одно фонетическое слово (информация о принадлежности слова к энклитикам или проклитикам была включена в список характеристик этого слова).

Как правило, все знаменательные части речи имеют одно сильное ударение. При этом некоторые из них могут иметь также одно или несколько слабых ударений, например: *ра=диолокацио+нный*[1], *мо=тове=лозаво+д* и т. п.

В исходном грамматическом словаре все слова помечены как полноударные, поэтому наличие и тип ударений в служебных частях речи корректировались вручную. Так, большинство предлогов и частиц в слитной речи являются либо слабоударными (как, например, предлог *ис-*

[1]Здесь и далее сильное ударение обозначается символом «+» после ударного гласного, слабое ударение – символом «=» после ударного гласного.

ходя=из, частица *абы=* и т. п) либо безударными (предлог *без*, частица *ведь* и т. п). В то же время существуют и сильноударные частицы (например, *почти+*, *пожа+луйста* и т. п). Кроме того, полно- либо частичноударными могут быть союзы (например, *наконе+ц* либо *бу=дто*).

Разработанный словарь, который включает в себя словоформы и их ЛГК, достаточно объемен. С целью более удобного использования словаря были сформированы теги. Теги представляют собой набор заглавных букв латинского алфавита, каждая из которых обозначает ЛГК словоформы, основываясь на ее английском названии (например, *инфинитив – I (Infinitive)*, *энклитика – E (Enclitic)*, *глагол – V (Verb)*). На рис. 2 используются следующие аббревиатуры (слева направо): N (Noun) – существительное, C (Common) – нарицательное, P (Proper) – собственное, A (Animated) – одушевленное, I (Inanimated) – неодушевленное, M (Masculine_gender) – мужской род, F (Feminine_gender) – женский род, C (Common_gender) – средний род, U (Undefined) – значение не выбрано (для множественного числа), S (Singular) – единственное число, P (Plural) – множественное число, N (Nominative_case) – именительный падеж (п.), G (Genitive_case) – родительный п., D (Dative_case) – дательный п., A (Accusative_case) – винительный п., I (Instrumental_case) – творительный п., P (Prepositional_case) – предложный п.

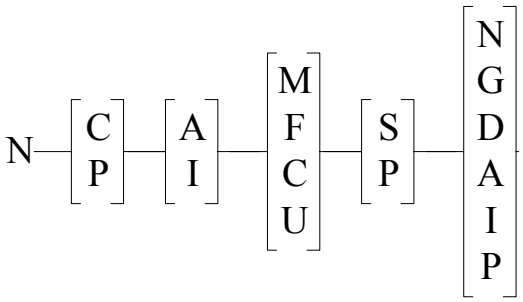

Рис. 2. Схема создания тегов для существительных

Из рис. 2 видно, что внутри каждой категории возможные значения уникальны (это является основным требованием к создаваемым тегам), однако бывают случаи, когда при использовании первых букв английского названия ЛГК происходит коллизия (например, в ЛГК времени глаголов *настоящее – P(Present)* и *прошедшее – P(Past)*). В таких ситуациях в одном из вариантов используется иная буква английского названия ЛГК (например, *настоящее время – S (preSent)*).

Наряду с уменьшением объема словаря использование тегов обеспечивает возможность выделения более общих групп частей речи (например, все существительные женского рода будут иметь тег NxxFxx, где x – любая буква (см. рис. 2).

Необходимо отметить, что при обработке словаря генерировались все словоформы, в том числе и потенциально возможные, которые практически никогда не встречаются в речи, но при необходимости все же могут быть образованы по общим правилам русского словоизменения, например краткая форма прилагательного *дереве+нский – дереве+нск* или страдательное причастие прошедшего времени от глагола *жда+ть – жда+нный*. Наличие в парадигме таких словоформ сопровождалось дополнительными пометами в словарной статье. При создании как можно более полного словаря такие пометы игнорировались. Однако при необходимости уменьшения объема лингвистических ресурсов (например, при создании системы синтеза речи для мобильных устройств) потенциально возможные словоформы будут первыми удалены из словаря.

Программные модули по формированию парадигм слов русского языка написаны на языке программирования C++. В качестве входного выступает текстовый файл, представляющий собой электронный вариант «Грамматического словаря русского языка» А.А. Зализняка и содержащий около 100 000 словарных статей.

На выходе программы формируются парадигмы слов каждой части речи. В некоторых парадигмах словоформы омонимичны между собой (например, и. п. – *кофе*; р. п., д. п., в. п.,

т. п., п. п. – *кофе*). Такие парадигмы сохраняются полностью, так как даже при наличии омонимии словоформ важную роль для синтеза речи играют их ЛГК.

В результате работы программы сформирован грамматический словарь, содержащий около 4 млн словоформ. Распределение количества элементов словаря по частям речи приведено в табл. 1.

Общее количество лексем в сформированном словаре – более 200 000, в то время как количество парадигм в исходном грамматическом словаре – менее 100 000. Такое увеличение количества парадигм произошло из-за того, что причастия и деепричастия были выделены в отдельные части речи, кроме того, глаголы, имеющие одновременно характеристики совершенного и несовершенного вида, представлены как две отдельные парадигмы.

Таблица 1

Количество слов и словоформ в базе, сформированной на основе «Грамматического словаря русского языка» А.А. Зализняка

Часть речи	Количество лексем в словаре	Количество словоформ в парадигме		Всего словоформ
		минимальное	максимальное	
Союз	564	1		564
Предлог	283	1		283
Частица	355	1		355
Междометие	184	1		184
Предикат	267	1		267
Вводное слово	65	1		65
Наречие	1372	1		1372
Местоимение	1076	1		1076
Существительное	44 747	6	13	545 621
Глагол	33 537	13	16	435 727
Причастие	69 018	27	32	1 931 537
Деепричастие	33 506	1	4	71 768
Прилагательное	16 006	27	58	909 023
Числительное	99	2	27	1473
Всего:	201 079	–		3 899 315

2. Пополнение словаря

В ходе создания словаря может возникнуть ряд проблем (ошибки в электронном словаре А.А. Зализняка, единичные отклонения от правил построения парадигм и т. п.), не позволяющих корректно сгенерировать все словоформы парадигмы. К тому же необходимо учитывать, что какой бы ни был большой словарь, в тексте вполне может встретиться слово, которое в нем отсутствует.

Представляется интересным получить некоторую оценку полноты сформированного словаря. Для такой оценки был выбран подкорпус Национального корпуса русского языка [7], доступный для свободного использования и представляющий собой случайную выборку предложений из корпуса со снятой омонимией объемом 180 000 словоупотреблений (90 000 – из прессы, по 30 000 из художественных, законодательных и научных текстов). Конечно, такой небольшой объем корпуса не позволяет точно оценить степень покрытия словарем текстов различных жанров, но дает возможность получить некоторые предварительные оценки, на основании которых можно принять решение о том, нужно ли создавать систему пополнения словаря.

В табл. 2, полученной на основе корпуса [7], представлены статистические данные о количестве слов из текстов различных жанров, которые отсутствуют в разработанном словаре. Количество отсутствующих в словаре слов показывает, насколько словарь нуждается в пополнении, в то время как количество отсутствующих в словаре слов с учетом их повторяемости – насколько важны добавляемые слова.

Таблица 2

Статистика отсутствующих в разработанном словаре слов

Тип текстов	Количество словоупотреблений в корпусе	Отсутствующие в словаре слова, %	Отсутствующие в словаре слова с учетом их повторяемости в корпусе, %
Художественные	30 003	6,4	9,1
Законодательные	30 004	3,3	12,9
Публицистические	90 018	7,3	13,8
Научные	30 002	9,3	17,7

Исходя из приведенной в таблице статистики, можно сделать вывод о необходимости разработки системы пополнения словаря.

2.1. Алгоритмы пополнения словаря

Процесс пополнения словаря можно разделить на следующие этапы:

– обнаружение в тексте слова, которого нет в словаре;

– добавление в словарь нового слова и всей его парадигмы.

На втором этапе происходит решение следующих задач:

– выделение в слове основы (стемминг);

– определение части речи и других несловоизменительных ЛГК;

– формирование всей парадигмы, к которой относится найденное слово, с расстановкой позиций ударений во всех словоформах;

– поиск сгенерированных словоформ в списке слов, не найденных в словаре, и их удаление из этого списка;

– сохранение сгенерированной парадигмы в словарь.

Существуют алгоритмы, позволяющие автоматически выделять основу слова и определять его ЛГК. Один из наиболее известных алгоритмов этого типа – стеммер Портера [8]. Суть алгоритма заключается в постепенном отбрасывании окончаний и суффиксов в слове с опорой на правила словообразования конкретного языка.

Отличительной особенностью алгоритма является то, что он не использует словари и базы основ слов. Алгоритм был адаптирован Портером для многих индоевропейских языков, в том числе и для русского. Тем не менее существуют особенности русского языка, не учтенные данным алгоритмом, что порождает ряд его недостатков, а именно:

– наличие похожих суффиксов, например *-ик-* и *-чик-*, неправильное определение которых приводит к некорректному выделению основы (например, *мячик*: здесь *«ч»* – часть корня, а *«ик»* – суффикс; *разведчик*: здесь *«чик»* – суффикс);

– наличие в языке слов-омографов[2] (например, множественное число, и. п. – *но+чи* и единственное число, р. п. – *ночи+*), что затрудняет определение ЛГК слова;

– наличие несклоняемых слов (таких как *кофе*, *хобби*, *бра*), при обработке которых также могут быть ошибочно определены их ЛГК.

Кроме того, стеммер Портера не предназначен для решения задач определения несловоизменительных ЛГК слов, а также для формирования всей парадигмы по найденной основе.

Известен другой алгоритм нахождения основы слова и его ЛГК – алгоритм определения грамматических характеристик словоформы методом графов [9]. Его особенности состоят в следующем:

– алгоритм основан на морфологическом анализе слова, при котором определяются суффиксы и окончания слов;

– по последнему суффиксу определяется принадлежность слова к той или иной части речи;

– по окончанию слова определяются остальные ЛГК с использованием заранее созданных для каждой части речи графов, в которых вершины – это некоторый набор ЛГК, а ребра – возможные окончания рассматриваемого слова.

[2]Омографы – слова, которые совпадают в написании, но различаются в произношении.

В сравнении со стеммером Портера данный алгоритм обладает следующими достоинствами:

– учитывается возможность наличия в словах русского языка более одного суффикса;

– учитывается возможность омонимии морфем, которая реализуется путем проведения дальнейшего анализа словоформы с помощью заранее созданного подграфа суффиксов для каждой части речи;

– существует возможность построения полной парадигмы на основании одной словоформы за счет использования графа наборов ЛГК и возможных окончаний для каждой части речи.

К недостаткам данного алгоритма можно отнести:

– необходимость наличия лингвистической базы данных, отражающей особенности словообразования и словоизменения для русского языка. Создание такой базы предполагает предварительную работу эксперта-лингвиста;

– невозможность определения словоизменительных ЛГК слов-омографов и несклоняемых существительных;

– невозможность определения несловоизменительных ЛГК.

Таким образом, существующие алгоритмы не дают возможности автоматически пополнять грамматический словарь. Исходя из этого, было принято решение о пополнении словаря вручную опытными экспертами-лингвистами с максимально возможной автоматизацией процедуры пополнения. Для этого был разработан алгоритм, блок-схема которого представлена на рис. 3. Основные действия, производимые экспертом-лингвистом вручную и программой автоматически, выделены соответствующими блоками. Последовательность действий показана номерами.

В соответствии с перечисленными выше этапами алгоритм включает поиск слов, отсутствующих в словаре (см. рис. 3, блоки 1–3), и занесение их в словарь (блоки 4–13). После выбора экспертом одного из временных слов и указания его части речи это слово автоматически помещается в поле для ввода неизменяемой части слова. При этом если в парадигме выбранной части речи более одной словоформы (состав парадигмы визуально представляется пользователю после указания части речи), то неизменяемая часть слова автоматически устанавливается во всех словоформах. Таким образом, любое изменение основы влечет за собой изменение всех словоформ. Извлеченные из БД наборы окончаний (извлечение происходит после указания несловоизменительных ЛГК слова) представляются для пользователя в виде списков возможных окончаний для каждой словоформы. Это позволяет эксперту выбрать одно из предложенных окончаний или указать иное путем непосредственного ввода. После выбора окончания какой-либо словоформы происходит автоматическое удаление наборов окончаний, которые в обрабатываемой словоформе не совпадают с окончанием, указанным экспертом. После генерации всех словоформ эксперт может сохранить парадигму. При этом автоматически генерируются два запроса в БД: один для сохранения введенной парадигмы, а второй для сохранения набора окончаний, использованного при ее построении. После успешного сохранения парадигмы автоматически организуется поиск сгенерированных словоформ в списке временных слов. Если какая-либо словоформа была найдена в списке, происходит ее удаление с последующим обновлением списка временных слов.

Разработанный алгоритм пополнения словаря обладает следующими достоинствами по сравнению с описанными выше алгоритмами:

– исключаются ошибки, связанные с вероятностным определением ЛГК и вероятностным выделением основы слова;

– решается задача определения несловоизменительных ЛГК;

– формирование парадигм происходит на основе особенностей построения каждой парадигмы в отдельности, а не на основе вероятностных наборов окончаний.

Кроме того, описанный автоматизированный подход значительно упрощает работу эксперта-лингвиста за счет реализации следующих функций:

– написание неизменяемой части слова (общей части для всех словоформ) требуется только один раз, во всех словоформах она отображается автоматически;

– процесс написания окончаний всех словоформ сводится к выбору одного из имеющихся в БД наборов окончаний (при отсутствии необходимого набора предоставляется возможность ввода окончаний вручную).

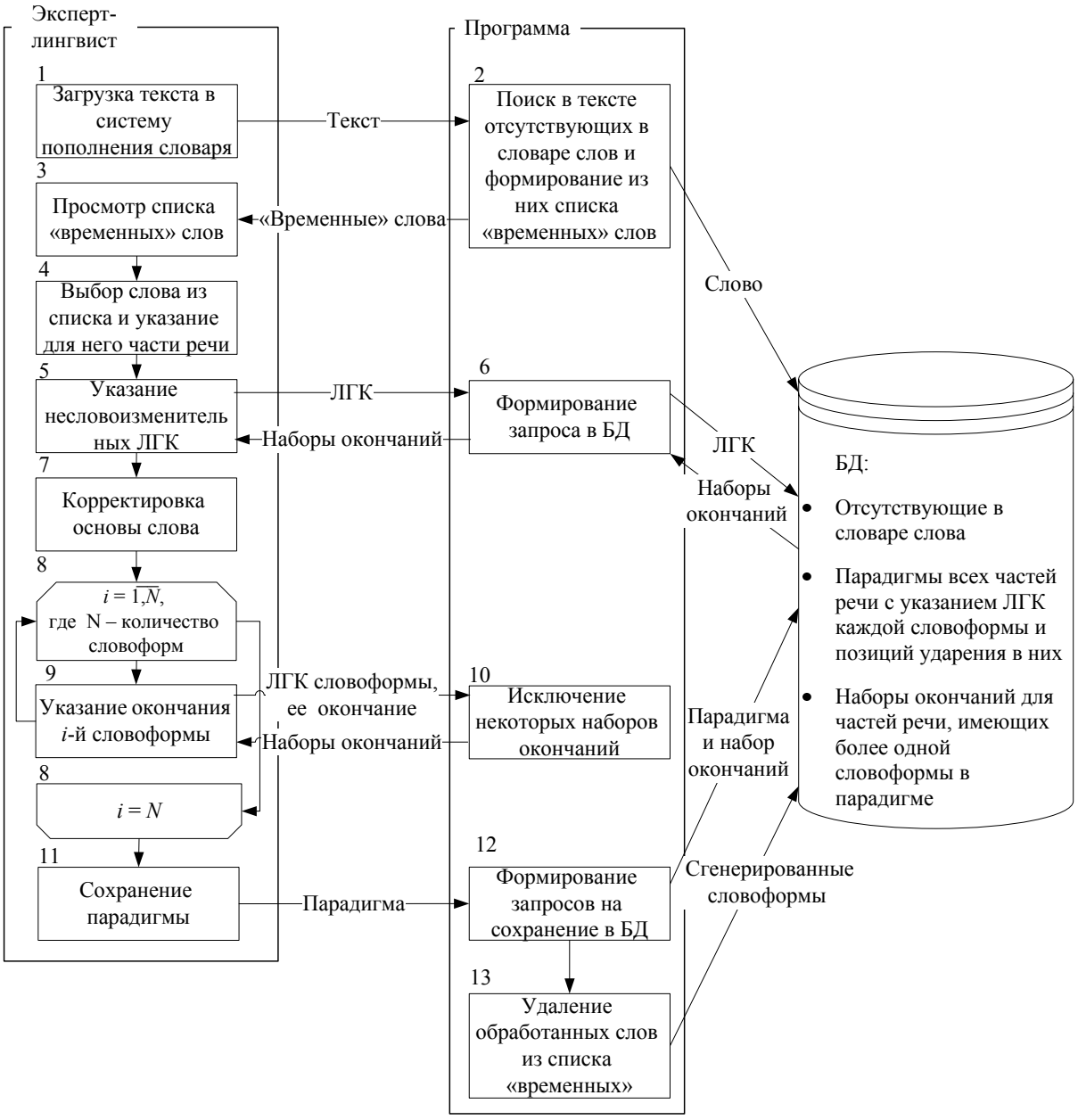

Рис. 3. Процесс автоматизированного пополнения словаря

2.2. Особенности программной реализации алгоритма пополнения словаря

Для реализации алгоритма пополнения словаря было разработано веб-приложение, использующее программный каркас Ruby on Rails [10], который основан на архитектуре Model-View-Controller (модель-представление-контроллер). Приложение взаимодействует с базой данных MySQL, где хранится словарь.

Использование веб-приложения дает возможность доступа к словарю многим пользователям, что способствует эффективности пополнения словаря. Коллизии, возможные при одновременном доступе к одному и тому же слову нескольких пользователей, разрешаются с помощью оптимистической блокировки [11], когда при одновременном редактировании одних и тех же данных несколькими пользователями первая их модификация сохраняется в БД, а предоставивший второе изменение пользователь при попытке сохранения получает сообщение об ошибке. Пользователь должен перезапустить веб-страницу, получая обновленные данные.

Доступ к функциям программы осуществляется после авторизации эксперта-лингвиста, пополняющего словарь. Главное меню включает в себя следующие пункты:

«Импорт» – функция, предоставляющая возможность эксперту загрузить в систему произвольный текст, в котором необходимо выбрать слова, отсутствующие в словаре;

«Временные слова» – функция отображения временных слов (слов из импортированного текста, отсутствующих в словаре) в виде списка;

«Обработанные части речи» – функция, предоставляющая возможность просмотра всех имеющихся в словаре парадигм в виде списка. Все парадигмы группируются по частям речи, к которым они относятся;

«Добавить слово» – функция, предоставляющая возможность добавления парадигмы какой-либо части речи, не основываясь на списке временных слов.

Разработанное приложение обладает следующими основными возможностями:

1. Для каждой части речи предусмотрено наличие некоторых особенностей. Например, для существительного – это возможность отсутствия в парадигме единственного или множественного числа, а также существования нескольких вариантов какой-либо словоформы. На рис. 4 отображена парадигма существительного *год*, в которой имеются по два варианта словоформ в именительном и винительном падежах множественного числа.

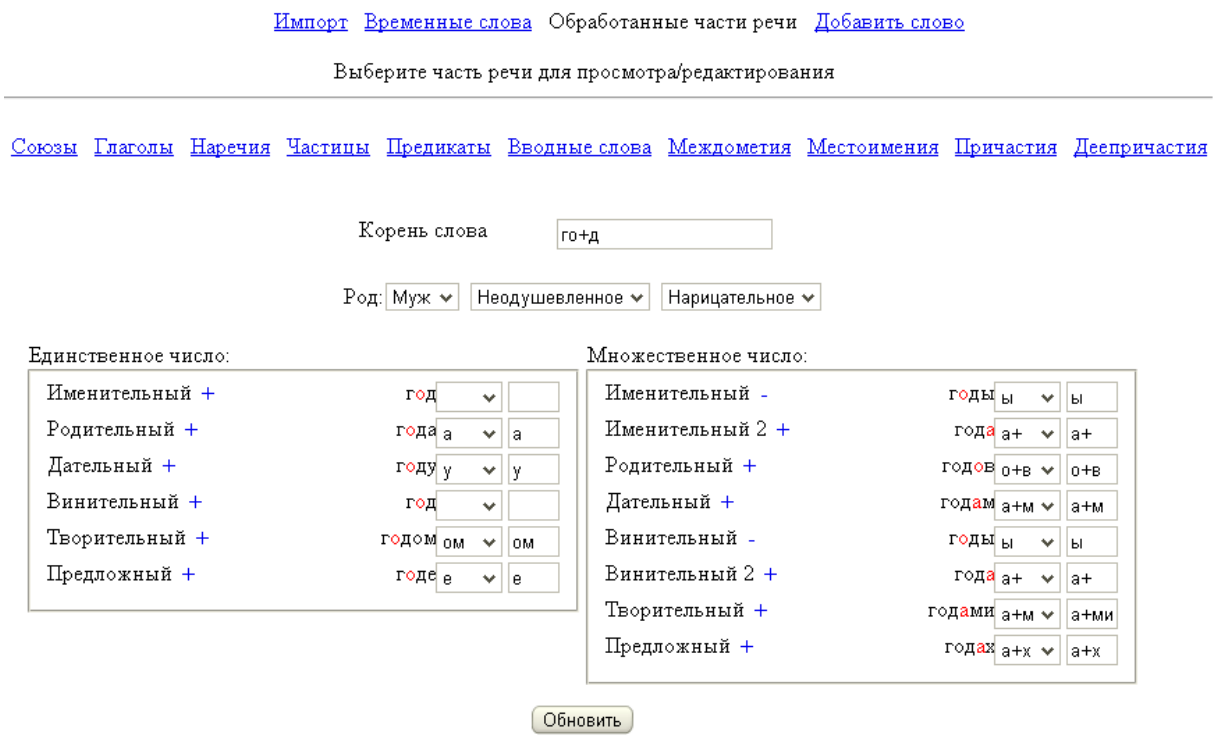

Рис. 4. Возможность введения дополнительных словоформ в парадигму

2. Можно отследить, содержится ли уже в словаре парадигма с такой же начальной формой. Если парадигма была найдена, то пользователю предлагается сравнить ее с той, которая только что была введена. При необходимости можно редактировать парадигму из словаря, удалять ее или добавлять новую парадигму в словарь в дополнение к имеющейся. На рис. 5 представлена довольно распространенная особенность русского языка, а именно двухвидовые глаголы, относящиеся одновременно к совершенному и несовершенному видам.

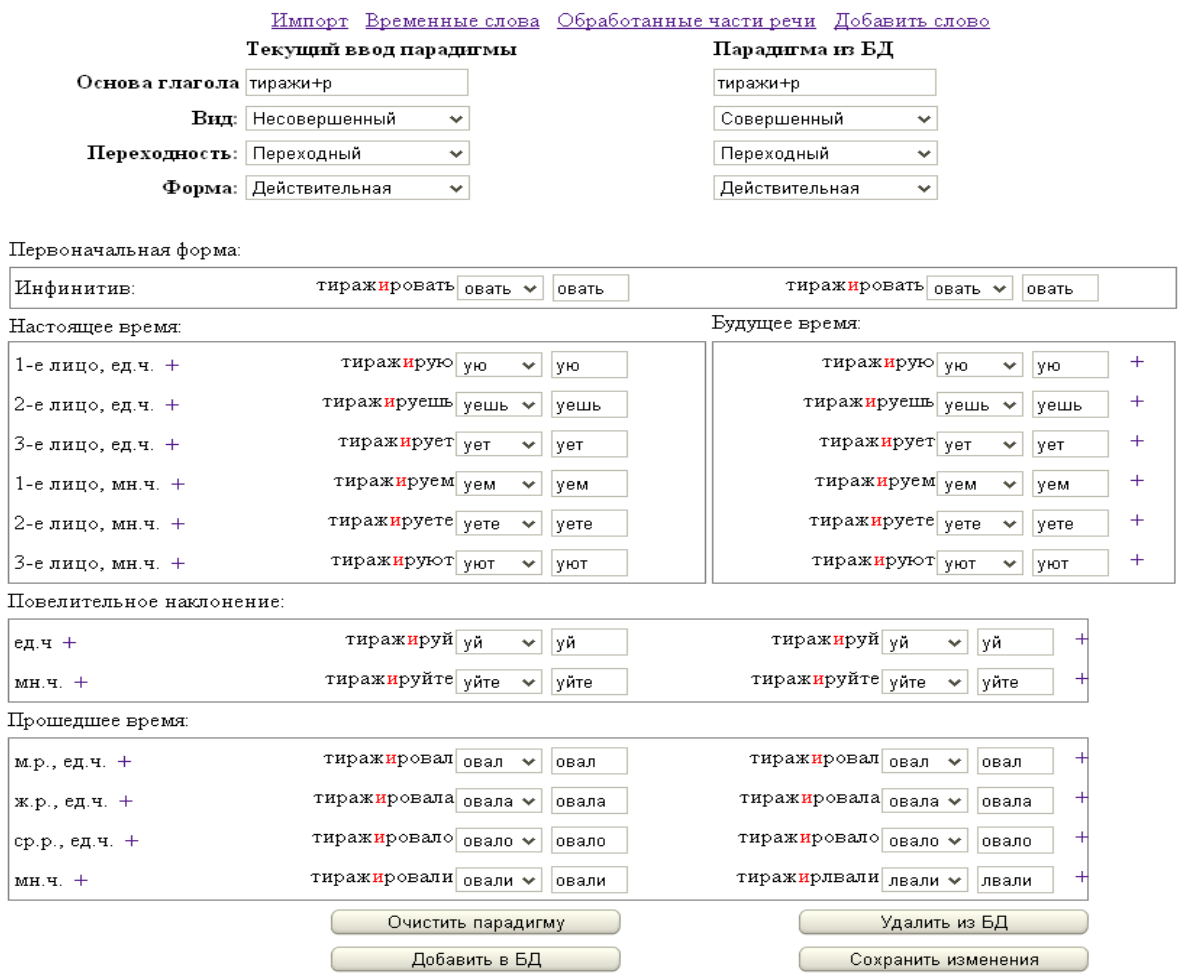

Рис. 5. Сравнение парадигм глаголов

Заключение

На основе разработанных алгоритмов создан электронный грамматический словарь русского языка, включающий около 4 млн словоформ (с учетом различных типов омонимии) и содержащий характеристики словесных ударений и лексико-грамматических категорий, необходимые для синтеза речи по тексту. Статистические исследования полноты словаря показали, что созданный словарь необходимо пополнять. Разработан и программно реализован алгоритм автоматизированного пополнения словаря, позволяющий добавлять в словарь всю парадигму слова и учитывающий особенности содержимого словаря для синтеза речи по тексту.

Список литературы

1. Speech synthesis ; ed. : J.L. Flanagan, L.R. Rabiner. – Dowden, Hutchinson & Ross, 1973. – 511 p.

2. Klatt, D.H. Review of text-to-speech conversion for English / D.H. Klatt // J. Acoust. Soc. Am. – 1987. – Vol. 82, № 3. – P. 737–793.

3. Dutoit, T. An Introduction to text-to-speech synthesis / T. Dutoit. – Kluwer Academic Publishers, 1997. – 286 p.

4. Lobanov, B. Development of multi-voice and multi-language TTS synthesizer (languages: Belarussian, Polish, Russian) / B. Lobanov, L. Tsirulnik // Speech and Computer : proc. of the 11th Intern. conf. SPECOM'2006, St. Petersburg, Russia, 25–29 June, 2006 / Institute of Informatics and Automation of RAS, Speech Informatics Group. – St.-Petersburg : Anatolia, 2006. – P. 274–283.

5. Зализняк, А.А. Грамматический словарь русского языка / А.А. Зализняк [Электронный ресурс]. – Режим доступа : http://starling.rinet.ru/cgi-bin/response.cgi?root=/usr/local/share/starling/morpho&morpho= 1&basename=\usr\local\share\starling\ morpho\zaliznia\dict&first=1. – Дата доступа : 17.06.201.

6. Зализняк, А.А. Грамматический словарь русского языка / А.А. Зализняк. – 2-е изд. – М. : Рус. яз., 1980. – 880 с.

7. Национальный корпус русского языка [Электронный ресурс]. – 2003. – Режим доступа : http://www.ruscorpora.ru. – Дата доступа : 17.06.2011.

8. Porter, M.F. An algorithm for suffix stripping / M.F. Porter // Program. – 1980. – Т. 14, № 3. – С. 130–137.

9. Вороной, С.М. Определение грамматических характеристик словоформы методом графов / С.М. Вороной, А.А. Егошина // Искусственный интеллект – 2008. – № 1. – С. 80–84.

10. Ruby on Rails [Electronic resource]. – 2004. – Mode of access : http://rubyonrails.org. – Date of access : 17.06.2011.

11. Коннолли, Т. Базы данных. Проектирование, реализация и сопровождение. Теория и практика : пер. с англ. / Т. Коннолли, К. Бегг ; под ред. К.А. Птицына. – 3-е изд. – М. : Вильямс, 2003. – 1440 с.

Поступила 15.12.11

[1]*Объединенный институт проблем информатики НАН Беларуси, Минск, Сурганова, 6 e-mail: liliya.tsirulnik@gmail.com*

[2]*Белорусский национальный технический университет, Минск, пр-т Независимости, 65 e-mail: veremei.vera@gmail.com*

L.I. Tsirulnik, V.V. Veremei

ALGORITHMS OF CREATION AND EXTENSION OF GRAMMATICAL DICTIONARY OF RUSSIAN LANGUAGE FOR TEXT-TO-SPEECH SYNTHESIS

Procedures of grammatical dictionary creation and extension are described. The peculiarities of word stress and additional grammatical characteristics which are used for TTS-synthesis are shown. An experiment for evaluating how well the dictionary covers texts of various genres has been conducted. The algorithm of automated extension of the dictionary is also proposed.