ДВУХУРОВНЕВЫЙ МОРФОФОНЕМНЫЙ ПРЕФИКСНЫЙ ГРАФ ДЛЯ ДЕКОДИРОВАНИЯ РУССКОЙ СЛИТНОЙ РЕЧИ

А. Л. Ронжин, Ан. Б. Леонтьева, И. А. Кагиров, Ал. Б. Леонтьева

Санкт-Петербургский институт информатики и автоматизации РАН

СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178

<{ronzhin,an_leo,kagirov,leonty}@iias.spb.su>

УДК 681.3

Ронжин А. Л., Леонтьева Ан. Б., Кагиров И. А., Леонтьева Ал. Б. **Двухуровневый морфофонемный префиксный граф для декодирования русской слитной речи** // Труды СПИИРАН. Вып. 4. — СПб.: Наука, 2007.

Аннотация. *Описан новый способ компактного хранения словаря слов и их транскрипций в виде фонемного графа, учитывающего дифференциальные морфологические признаки слов. Сокращение словаря особенно актуально для флективных языков, где богатая морфология сильно затрудняет анализ текста и речи. Для повышения производительности декодера русской речи с большим словарем предлагается использовать двухуровневый морфофонемный префиксный граф. Выделение одинаковых основ и окончаний в различных словах существенно сокращает пространство поиска гипотез распознавания. Использованная статистическая модель языка учитывает встречаемость комбинаций основ, а не целых слов, что уменьшает сложность декодирования слитной речи и требует для обучения значительно меньшего объема текстовых ресурсов. По сравнению с базовыми моделями фонетического представления словаря сложность топологии предложенного графа оказалась в 17 раз меньше.* — Библ. 12 назв.

UDC 681.3

Ronzhin A. L., Leontyeva An. B., Kagirov I. A., Leontyeva Al. B. **Two-level morpho-phonetic prefix graph for the Russian continuous speech decoding** // SPIIRAS Proceedings. Issue 4. — SPb.: Nauka, 2007.

Abstract. *A new representation structure of large vocabulary for high inflective language is sketched. Reach morphology complicates text and speech parsing. To improve the performance a two level morpho-phonetic prefix graph is proposed for vocabulary representation. Sharing the identical beginning parts and endings of different words significantly reduces the search space for a large vocabulary. Stem based language model reduces the complexity of continuous speech decoding and solves data scarcity problem for the inflective languages. The proposed graph was compared with two baseline word lattice models that showed significant reduction of topology complexity of the graph.* — Bibl. 12 items.

1. Введение*

Модели распознавания речи флективных и агглютинативных языков активно изучаются в последнее время. Причина в том, что множество европейских языков, таких как финский, немецкий, литовский, эстонский, турецкий и множество славянских, имеют общие особенности. Однако пословный декодер речи, который используется в большинстве современных систем распознавания речи и достаточно успешно подходит для английского, не годится для флективных языков по скорости и точности вследствие их богатой морфологии. Вышеперечисленные языки обладают развитыми системами словообразования и словоизменения, что автоматически увеличивает их словарь в несколько раз по сравнению с языками аналитического строя [1]. Для уменьшения размера

* Данное исследование проводится в рамках Европейской научной сети SIMILAR NoE FP6-IST-2002#507609, проектов INTAS № 04-77-7404 и № 05-1000007-426, а также гранта Президента Российской Федерации № МК-9351.2006.9 и гранта РФФИ № 07-07-00073-а.

словаря, упрощения языковой модели и сокращения времени декодирования вместо целых слов были исследованы различные единицы, представляющие части слов [2].

Различия между существующими подходами заключаются в типе используемых единиц и способе синтеза и разложения слова при распознавании и обучении соответственно [3]. Здесь можно выделить два основных подхода: методы сегментации, основанные на грамматических правилах, и декомпозиция слов на морфемы статистическим методом [2]. При разложении слова на составляющие единицы неизбежно приходится решать вопросы, связанные с семантической неоднозначностью. Первый подход может обеспечить это разделение, но требует громадного ручного труда и аккуратного программирования [1,4], в то время как второй, используя метод обучения без учителя, автоматически находит морфемоподобные единицы (статистические морфы). Алгоритм декомпозиции слов на морфы статистическим методом Morfessor, представленный в [2], был успешно использован при распознавании финской, эстонской и турецкой речи.

В ходе применения морфемного декодера для флективных языков возникают трудности, связанные не только с моделью языка, но и с отсутствием речевых и текстовых данных для обучения. По этой причине морфемные модели, построенные по строгим грамматическим правилам, кажутся более перспективными, так как они более полно описывают все возможные словоформы и их комбинации в произвольном тексте, которые не могут наблюдаться в больших, но тем не менее ограниченных обучающих текстовых данных.

Введение морфемного уровня обработки речевого сигнала в основную систему распознавания речи требует решения некоторых проблем. В первую очередь следует отметить, что когда морфемы используются вместо целых слов, то существует дополнительная проблема, связанная с объединением морфов в слова или нахождением границ слов в потоке морфов [5]. В спонтанной речи паузы между словами обычно не очевидны, и поэтому достаточно трудно с акустической точки зрения выделить слова в распознанной последовательности морфов.

Для решения этой проблемы при статистическом морфемном подходе в процессе обучения языковой модели и декодирования речи используется специальный морф «конец слова» [2]. В грамматических морфемных моделях анализ типов соседних морфем позволяет определить границы слова. Однако некоторые морфемы могут принадлежать нескольким типам (приставка, корень, суффикс и т.д.) и в этом случае трудно определить границы слова. В процессе декодирования морфемы распознаются методом back-off, который обеспечивает формирование наиболее вероятной с акустической точки зрения последовательности морфем [4]. В декодере морфемная сеть содержит список цепочек контекстонезависимых или основных фонем, который формирует произношение каждого морфа в словаре. Поскольку в процессе распознавания используются не целые слова, то есть риск появления грамматически неправильных, но акустически очень близких к записанному речевому сигналу слов [2]. Особенно часто это случается, если морфы короткие. Благодаря использованию морфов система способна распознавать слова, которых нет в словаре, однако существует большая вероятность появления грамматически неправильных слов. Кроме того, морфы гораздо короче, чем целые слова, и поэтому чаще путаются, вследствие их большей акустической схожести.

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Стохастическая морфосинтаксическая языковая модель (СМЯМ) была предложена в [4], чтобы избежать грамматически некорректных слов и предложений. Эта модель использует морфосинтаксическую грамматику как автомат конечных состояний в комбинации с N-граммной моделью языка. Представление флективных словоформ СМЯМ включает в себя преимущества детерминистической грамматики и стохастического подхода. Морфосинтаксическая модель отфильтровывает грамматически неправильные комбинации, но при этом увеличивается размер сети. Таким образом, эта модель использует сильные ограничения на связь между различными классами морфов, в то время как стандартная N-граммная модель позволяет любые комбинации.

Принимая во внимания преимущества морфемного декодера речи и компактного представления словаря как фонетического префиксного дерева [6], был разработан двухуровневый морфофонемный префиксный граф (ДМПГ) для русского словаря и соответствующий декодер речи. Для увеличения эффективности поиска словарь организуется как древовидная структура: объединяются общие приставки и окончания слов, по мере возможности удаляются дублирующие пути [7]. Кроме того, применяется модель языка, построенная на основах, а не на целых словах, так как она более полно описывает языковые структуры и требует для обучения гораздо меньше текстовых ресурсов [8]. Чтобы адаптировать модель для распознавания в режиме реального времени, вычисление банка вероятностей всех фонем производится перед декодированием слов с использованием параллельных процессоров [9].

2. Морфологический подход к анализу русского языка

Для сокращения затрат времени на грамматическую интерпретацию входящего текста на русском языке имеет смысл использовать модуль морфологического анализа. Использование морфологического модуля позволяет избежать увеличения словаря: в словаре хранятся только основные формы лексем и нет отсылок к косвенным (падежные формы, личные формы глагола и т.п.), так как образование косвенных форм происходит по строгим словоизменительным правилам. Имея информацию об этих правилах, модуль способен самостоятельно образовать любую косвенную форму или произвести ее адекватную грамматическую атрибуцию.

Модуль морфологического анализа выделяет в словоформе лексическую основу и грамматические показатели. Все лексические основы принадлежат словарю, тогда как грамматические показатели, которых существенно меньше, чем лексических основ, образуют парадигмы по определенным правилам. Таким образом, целый ряд словоформ, различающихся только грамматическими показателями, трактуется как одна единица словаря.

Фактически, морфоанализатор отсеивает все регулярное (грамматику) на входе, оставляя словарю все нерегулярное, то есть лексические значения (лексемы). Кроме того, использование модуля морфологического анализа позволяет автоматически определять морфологические значения, выраженные на словоформах.

В 2006–2007 гг. группой речевой информатики СПИИРАН был создан программный модуль «Диаморф», представляющий собой реализацию описанной выше концепции. Работа над модулем является частью проекта по созданию системы распознавания слитной речи на русском языке. Предполагается использование морфоанализатора при создании систем автоматического синтеза

и распознавания речи на русском языке. Кроме того, процедура морфемного анализа необходима для создания морфостатистической модели языка, которая позволяла бы предсказывать вероятность сочетания тех или иных морфем в пределах одной словоформы. Подобное представление языка также может повысить эффективность систем автоматического распознавания речи. Следует отметить, что существуют альтернативные подходы к анализу русского языка [10,11], однако в настоящей статье они не рассматриваются.

В настоящей статье рассматривается структура программного модуля морфологического анализа только для имен существительных, прилагательных и некоторых местоимений.

2.1. Система грамматических правил

В русском языке существуют четкие правила, по которым к основе слова присоединяются окончания и грамматические суффиксы. В общем случае, правила соотносят определенные парадигмы суффиксов и/или окончаний и классы слов (традиционно именуемые типами спряжения или склонения). Эти правила словоизменения представлены в теоретической части «Грамматического словаря русского языка» А. А. Зализняка.

Каждый грамматический показатель $\mathbf{X} = \{x_1, x_2, ..., x_N\}$ имеет определенную грамматическую семантику (морфологические дифференциальные признаки) $\mathbf{G} = \{g_1, g_2, ..., g_K\}$, что можно представить в виде бинарной матрицы **P** размером $N \times K$, где

$$p_{ij} = \begin{cases} 1, \text{если признак } g_j \text{ выполняется} \\ 0, \text{иначе} \end{cases}.$$

Т.е. элементы матрицы **P** будут равны единице только в том случае, если морфологический дифференциальный признак g_j и конкретный грамматический показатель x_i соответствуют друг другу.

В качестве примера парадигму слова «стол» представим в виде таблицы (табл. 1), однозначно соотносящей морфологические показатели (именные окончания «Ø», «-а», «-у», «-ом», «-е», «-ы», «-ов», «-ам», «-ами», «-ах») и присущие им значения $(g_1 \mathbf{K} g_k)$. Семантика каждого окончания определяется родом, числом и падежом. Парадигматика окончаний обеспечивается такими характеристиками, как тип и номер склонения.

В табл. 1 использованы следующие обозначения: С — субстантивное склонение, А — адъективное склонение, М — местоименное склонение; цифрой обозначен номер склонения; Е и М в графе «Число» — единственное и множественное число соответственно; М, Ж, С в графе «Род» — мужской, средний и женский род соответственно; буквы в графе «Падеж» — сокращения от названий падежей русского языка: ВО — винительный падеж одушевленных имен сущ.; ТУ — устаревшая форма творительного падежа: рукой (норм., Т) — рукою (устар., ТУ); КФ — краткая форма имен прилагательных; признак шипящей — проверка на наличие шипящего перед окончанием: «надо» — 1 в случае необходимости проверки, иначе 0; «есть» — наличие шипящего перед окончанием. Знаком «+» обозначается нулевое окончание (Ø).

Для определения полной парадигмы лексемы нужно выбрать все окончания по графе падеж, которые удовлетворяют требованиям граф Тип склонения, Номер склонения, Род. В некоторых случаях, особо помеченных в словаре, вы-

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

бираются окончания из графы «Особые окончания», а также дополнительно учитываются ударность и наличие шипящей.

Таблица 1

Формальное представление парадигмы лексемы «стол»

		+	а	у	+	ом	е	ы	ов	ам	ы	ами	ах
Тип склонения	С	1	1	1	1	1	1	1	1	1	1	1	1
	А	0	0	0	0	0	0	0	0	0	0	0	0
	М	0	0	0	0	0	0	0	0	0	0	0	0
Номер склонения	1	1	1	1	1	1	1	1	1	1	1	1	1
	2	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0	0	0	0
Число	Е	1	1	1	1	1	1	0	0	0	0	0	0
	М	0	0	0	0	0	0	1	1	1	1	1	1
Род	М	1	1	1	1	1	1	1	1	1	1	1	1
	Ж	0	0	0	0	0	0	0	0	0	0	0	0
	С	0	0	0	0	0	0	0	0	0	0	0	0
Падеж	И	1	0	0	0	0	0	1	0	0	0	0	0
	Р	0	1	0	0	0	0	0	1	0	0	0	0
	Д	0	0	1	0	0	0	0	0	1	0	0	0
	В	0	0	0	1	0	0	0	0	0	1	0	0
	ВО	0	0	0	0	0	0	0	0	0	0	0	0
	Т	0	0	0	0	1	0	0	0	0	0	1	0
	ТУ	0	0	0	0	0	0	0	0	0	0	0	0
	П	0	0	0	0	0	1	0	0	0	0	0	1
КФ		0	0	0	0	0	0	0	0	0	0	0	0
Шипящая	надо	0	0	0	0	0	0	0	0	0	0	0	0
	есть	0	0	0	0	0	0	0	0	0	0	0	0
«Неправильные» окончания	1°	0	0	0	0	0	0	0	0	0	0	0	0
	2°	0	0	0	0	0	0	0	0	0	0	0	0
	3°	0	0	0	0	0	0	0	0	0	0	0	0
Ударение	надо	0	0	0	0	0	0	0	0	0	0	0	0
	есть	0	0	0	0	0	0	0	0	0	0	0	0

Полный вариант таблицы включает в себя 1459 строк, 43 столбца и учитывает также слова, изменяющиеся по падежам особым образом. Список помет, часть которых использовалась при создании модуля «Диаморф», приводится в [1].

Отдельные формы каждой лексемы различаются не только суффиксами и окончаниями, но и основами. Различия в основах могут быть двух типов: различные словоформы образуются от разных основ, например идти–шел, или в основе происходят регулярные изменения: любовь–любви.

Все случаи первого типа признаются уникальными и заносятся в словарь в виде комментариев к статьям. Случаи второго типа описываются определенными правилами, что в словаре выражается при помощи помет у соответствующей лексемы.

Для модуля «Диаморф» была создана база данных по основам, в которой каждая лексема представлена одной или более основой, причем каждая основа

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

имеет валентность на определенные окончания. Например, основы лексемы «конец» представлены как:

- конец- (им. ед.ч.);
- конц- (косв.п. ед.ч.; мн.ч.).

Основы генерируются по правилам, описанным в [1], а именно:

1. В словах с пометами м, мо, ж (кроме 8*), жо (кроме 8*), мс, мс-п, отмеченных астериском (*), основы для всех косвенных падежей (кроме Тв.п. ед.ч. с окончанием -ью):
   а) с последним гласным «о»: чередование -о-/-Ø-;
   б) с последним гласным «и»: чередование -и-/-ь-;
   в) с последним гласным(V)+«е» чередование -е-/-й-;
   г) с последним гласным «е» в типе м (мо) 6* чередование -е-/-ь-;
   д) с последним согласным(С) (кроме «ш», «щ», «ж», «ч», «ц»)+«е» в типе м (мо) 3* чередование -е-/-ь-;
   е) с «л» перед последним V «е» меняется -е-/-ь-;
   ж) во всех прочих случаях, когда последний гласный основы «е» меняется -е-/-Ø-.
2. В словах с пометами ж (кроме 8*), жо (кроме 8*), с, со, мс, мс-п, отмеченных астериском, основы для Род.п. мн.ч. (и Вин.п. мн.ч., если он совп. с Род.п. мн.ч.):
   а) конечный -ь- основы в типах ж (жо) 6*, с (со) 6*, то чередование -ь-/-é- (-и-);
   б) конечный -ь-С или -й-С не в типах ж (жо) 6*, с (со) 6*, то -ь-С (-й-С) /е-С;
   в) -к/-г/-х-С/-к/-г/-х-о-С;
   г) тип 3: перед -к/-г/-х, но не после -ш, -ж, -ч, -щ, -ц ставится «о»;
   д) в остальных положениях ставится «е» (после шип. -о-).
3. В словах с пометами п • в основе для краткой формы мужского рода конечный -н- основы отпадает.
4. В словах с пометами п , в основе для краткой формы конечный -н- основы отпадает.

2.2. Алгоритм работы морфоанализатора

Общий принцип работы морфоанализатора состоит в том, что словоформа разбивается на лексическую основу и грамматические показатели, после чего ищется вся парадигма конкретной лексемы. Ниже, на рис. 1 представлена блок-схема алгоритма морфемного анализа. На вход программного модуля подается список слов в виде текстового файла. Каждая строка этого файла содержит одну словоформу в любом числе, падеже и роде (для прилагательных и местоимений). На выходе генерируется файл, в котором для каждого слова построена парадигма слова для всех падежей, единственного и множественного числа, если это возможно.

На этапе разработки программного модуля была сгенерирована база основ. Для ее составления использовался словарь А. А. Зализняка [1]. Она содержит основы слов с пометами. Также были сгенерированы дополнительные основы, образуемые в случае чередования в основе.

При анализе словоформы программный модуль вычленяет так называемое гипотетическое окончание. Происходит это следующим образом: сначала исходная словоформа сравнивается со всеми основами в базе данных, если

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

найдена основа, то окончание принимается нулевым и строится парадигма в соответствии с пометами основы. Далее исходная словоформа сопоставляется со словоформами парадигмы. Если найдено совпадение, это означает, что основа выбрана правильно и построенная парадигма печатается в файл. Если первичный поиск не дал результатов, то окончание считается ненулевым, от исходной словоформы отрезается последняя буква, которая сохраняется в символьном массиве, а что остается от словоформы объявляется гипотетической основой. Гипотетическому окончанию приписывается набор грамматических значений в соответствии с приведенной выше табл. 1. Далее гипотетическая основа вновь сопоставляется с основами из базы данных, имеющими пометы на присоединение грамматических показателей с определенными значениями. Если эти пометы и признаки гипотетического окончания совпадают, разбиение словоформы на основу и окончание считается верным. Длина гипотетического окончания составляет не более трех символов. Если в результате поиска не обнаружена основа, то в выходной файл выводится сообщение: «Основа не найдена».

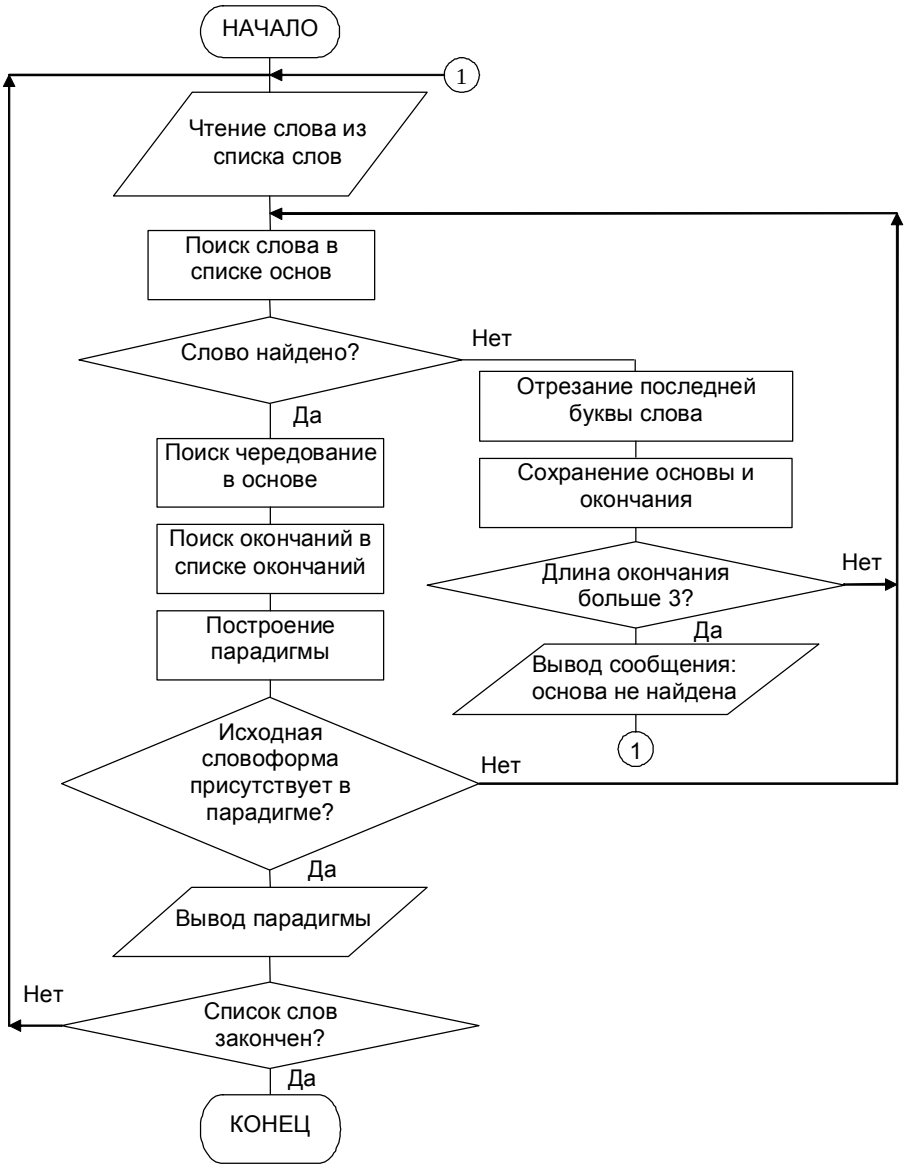


Рис. 1. Блок-схема алгоритма морфемного анализа.

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Построение парадигмы происходит следующим образом. Когда основа найдена, анализируются соответствующие ей пометы, указывающие на часть речи, тип и номер склонения, и другие признаки, определяющие правила построения парадигмы. У некоторых словоформ присутствует чередование в основе, на которое указывает помета «*», а для прилагательных — еще и пометы «1» или «2». Различают несколько типов чередования, в зависимости от которых при построении парадигмы учитываются одна или две дополнительные основы. После проверки на чередование в основе осуществляется поиск возможных окончаний в соответствии с таблицей признаков. Далее строится парадигма исходного слова, причем вид парадигмы определяется пометами при найденной основе. Парадигма сохраняется в файле в удобном для пользователя виде. Если не возможно полное или частичное построение парадигмы, то выводятся соответствующие сообщения.

Разработанный программный модуль был использован для обработки словаря А. А. Зализняка. Первоначальный объем словаря составляет 97194 слова. Была сгенерирована расширенная база данных, содержащая исходные основы и дополнительные основы, получаемые в результате чередования; ее объем составил 117351 основу. В качестве проверки был обработан корпус словаря А. А. Зализняка и получена соответствующая статистика. Например, число парадигм для существительных составило 621682 слова, а для прилагательных — 882102 слова. Если при распознавании речи использовать подход «фонемы–словоформы», то только для существительных объем базы данных составил бы более 600 тысяч слов, в то время как при морфологическом анализе количество основ существительных составляет около 46 тысяч. Таким образом, очевидно, что использование морфологического модуля позволяет существенно сократить объем словаря. На рис. 2 приведена диаграмма распределения словоформ по частям речи, на рис. 3 отображено распределение слов с различными видами типами чередования.

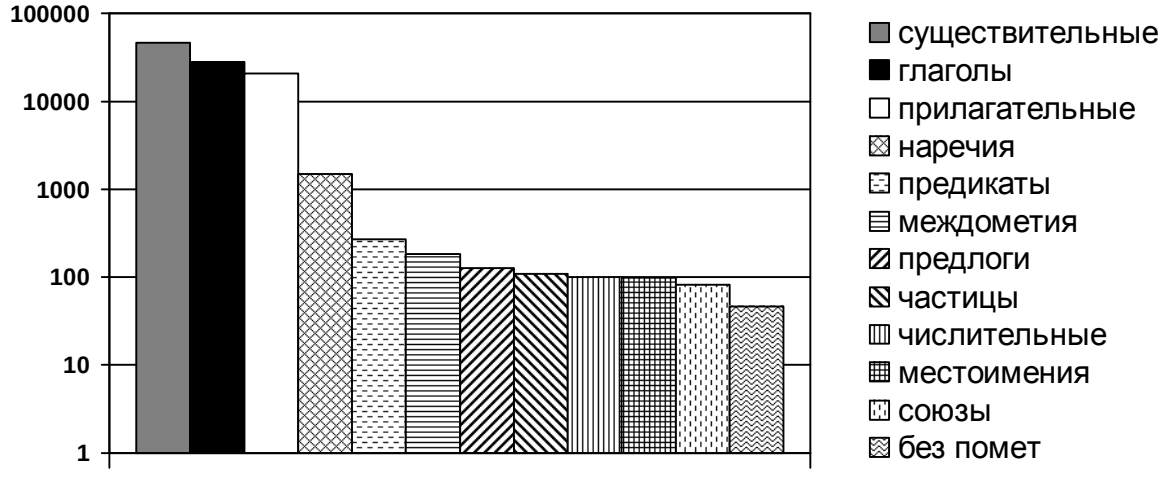


Рис. 2. Распределение словоформ по частям речи.

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

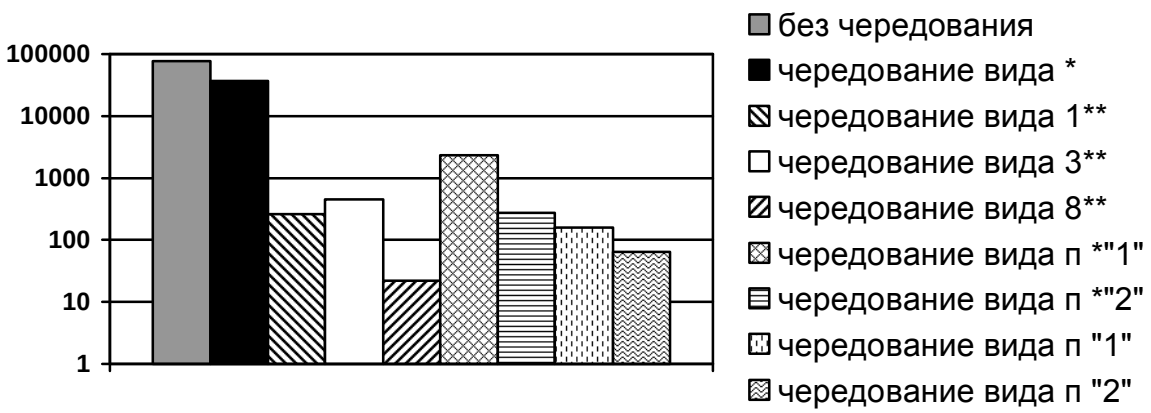


Рис. 3. Распределение словоформ по типам чередования в основе.

Из графиков видно, что глаголы по своей численности занимают второе место после существительных, однако количество словоформ, которые они порождают, значительно больше. Это дает существенный выигрыш в размерах базы данных. Учет чередования в основе несколько увеличил количество основ, так как довольно многие лексемы относятся к данному типу; более того, слова могут иметь одну или две дополнительные основы, в зависимости от типа чередования. Тем не менее подобное увеличение базы данных несущественно по сравнению с размером базы данных, содержащей все словоформы.

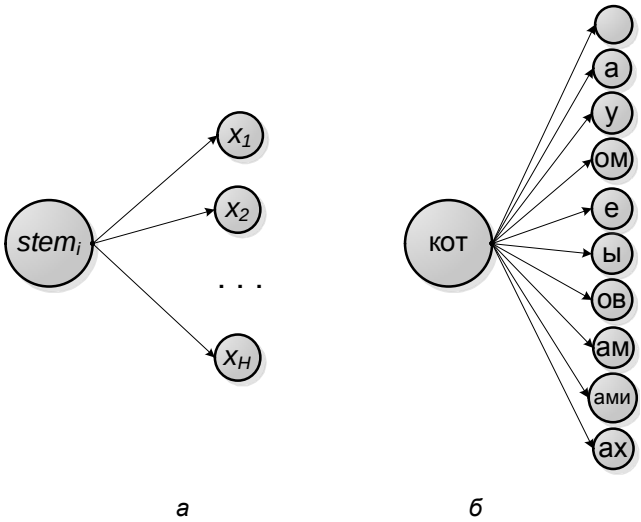
3. Представление парадигмы слова в виде морфофонемного графа

Для компактного хранения и быстрого доступа к набору словоформ, производных от одной основы, используем ориентированный граф, целиком описывающий парадигму слова. Узлами данного графа являются основы и окончания. Тогда при разбиении словоформы на основу и грамматические аффиксы структура ее графа будет состоять из некоторой основы слова $stem_i$, связанной с приемлемыми для нее грамматическими аффиксами $\{x_1, x_2, \mathbf{K}, x_H\}$ (рис. 4*а*). Большинство слов строится посредством такого графа. Следует отметить, что в структуру заносятся только неповторяющиеся окончания, однако на этапе синтаксического анализа одно и то же окончание может соответствовать нескольким грамматическим показателям. Так, в примере на рис. 4*б* окончание «а» служит для образования двух различных грамматических словоформ («кота» в родительном падеже и «кота» в винительном падеже).

В ряде случаев при формировании словоформ наблюдаются изменения не только в грамматических аффиксах, но и в самой основе (см. раздел 2). Чтобы учесть возможные варианты чередования в основе, в структуру графа вводятся несколько вариантов основы $\{stem_{i1}, stem_{i2}, \mathbf{K}, stem_{iN}\}$. Причем для каждого варианта основы существует свой набор грамматических аффиксов из множества **X** (рис. 5*а*). Если для разных вариантов основ встречается одинаковый аффикс, то производится объединение соответствующих узлов. В результате один и тот же аффикс может быть соединен с несколькими основами одновременно. Примеры такого графа приведены для слов «конец» (рис. 5*б*) и «идти» (рис. 5*в*). В первом примере возникает чередование второго типа с регулярным изменением в основе. Во втором примере присутствует как первый

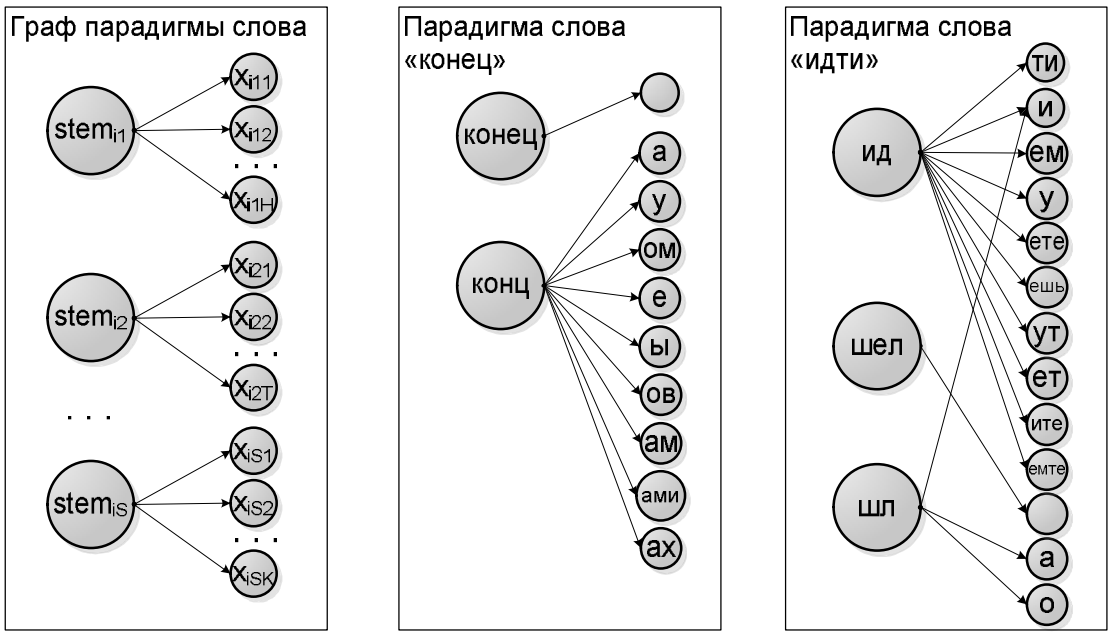
Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

тип чередования, где различные словоформы образуются от разных основ («идти» — «шел»), так и второй («шел» — «шли»). Представленный на рис. 5*в* граф образует только личные и неопределенную формы глагола, полный же граф включает в себя и остальные вербоиды (причастия и деепричастия). При этом число основ возрастает до пяти (добавляются «шедш» и «идущ»), а число аффиксов — до 26 (добавляются «ая», «его», «ее», «ей», «ем», «ему», «ею», «ие», «ий», «им», «ими», «их», «ую», «я»).



а *б*

Рис. 4. Ориентированный граф для представления парадигмы слова.
а — общая форма; *б* — пример графа для слова «кот».



а *б* *в*

Рис. 5. Граф парадигмы слова, учитывающий чередование в основе.
а — общая форма; *б* — пример графа для слова «конец»; *в* — пример графа для слова «идти».

Представленный в разделе 2 модуль «Диаморф» обеспечивает генерацию полной парадигмы для произвольного слова или основы из словаря. С его помощью формируются все словоформы словаря. Следующим этапом при создании базы данных для декодера речи является транскрибирование всех слово-

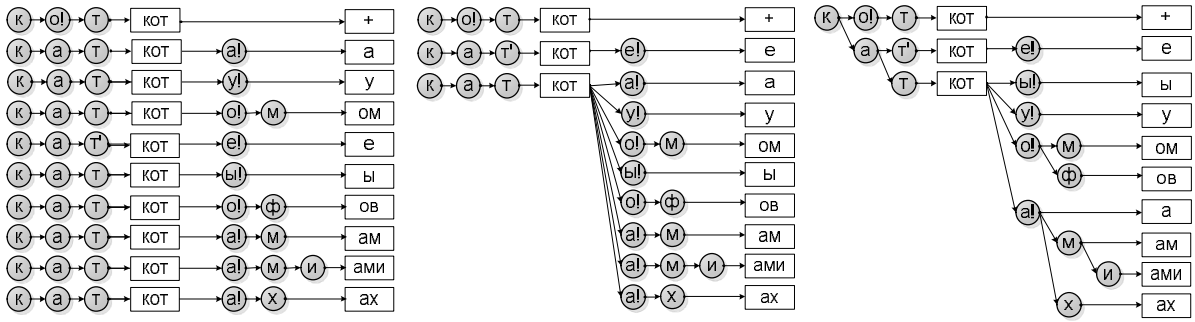
Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

форм с последующим представлением их в виде единого морфофонемного графа.

Используя правила транскрибирования, производится перевод всех возможных словоформ из графемного в фонетическое представление [5]. На рис. 6 представлены три варианта графов для транскрибированного представления парадигмы слова «кот». Узлами здесь являются фонемы, а также основы и окончания. На рис. 6*а* транскрипции 10 словоформ, представленных на рис. 4*б*, записаны в виде списка цепочек соответствующих фонем. Следует заметить, что в фонетической записи число вариантов основ и окончаний несколько возрастает. Например, в данном случае число основ увеличилось до трех («к-о!-т», «к-а-т», «к-а-т'») вследствие учета ударных и безударных гласных, а также мягких и твердых согласных. В разработанной модели автоматического транскрибирования текста всего используется 44 фонемы.

Теперь, объединяя одинаковые варианты транскрипций основ и окончаний, приведем список транскрибированных словоформ (рис. 6*а*) к более компактной записи в виде морфофонемного графа (рис. 6*б*). При этом полученный граф содержит транскрипцию основ и окончаний, а также хранит их графемные формы.

Далее производится объединение идентичных префиксных путей графа, начиная с первых фонем основ и с первых фонем окончаний, посредством следующей процедуры. Последовательно сравниваются первые фонемы всех основ и одинаковые объединяются в один узел графа. Затем подобная операция выполняется внутри каждого подграфа со вторыми фонемами всех основ и так далее, пока не будет достигнута последняя фонема самой длинной основы (рис. 6*в*). Следует учесть, что при обработке окончаний анализ ведется в рамках каждого варианта основы отдельно, поскольку объединение одинаковых первых фонем окончаний от разных вариантов основ может привести к генерации грамматически некорректной словоформы.



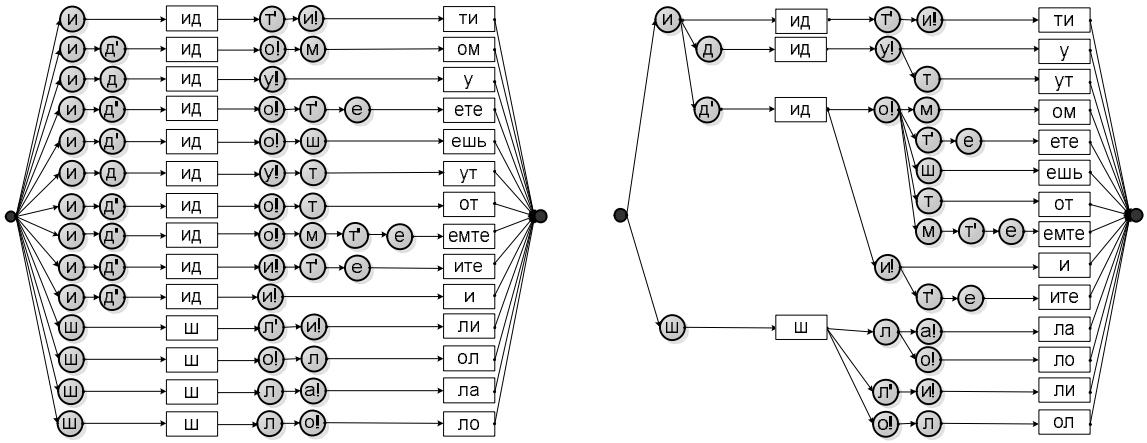
а *б* *в*

Рис. 6. Варианты представления транскрипций парадигмы слова «кот».
а — список; *б* — морфофонемный граф; *в* — морфофонемный префиксный граф.

Таким образом, парадигма слова и ее транскрипции представлены в виде двухуровневого морфофонемного префиксного графа (ДМПГ). Первый уровень графа служит для хранения регулярной части парадигмы в виде префиксного дерева, где корневым узлом является первая фонема слова, а листьями являются варианты основ. Второй уровень представляет собой лес префиксных деревьев и содержит все возможные для основ грамматические аффиксы. В общем случае, оба уровня представляют собой ориентированные графы, поскольку при чередовании в основе число начальных узлов может быть больше

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

одного. Кроме того, полностью одинаковые транскрипции окончаний могут встречаться у разных вариантов основ и тогда концевой узел, представляющий окончание, будет связан с несколькими узлами. Так, на рис. 7*б* граф, построенный для слова «идти», содержит 2 начальных узла, представляющих фонемы «и» и «ш». Как замечено ранее, объединение фонем окончаний производится в рамках каждого варианта основы независимо, поэтому в графе окончаний после сокращения узлов остались две фонемы «о!» во избежание генерации неправильных слов («шом», «шоте» и т.д.).



а *б*

Рис. 7. Варианты представления транскрипций парадигмы слова «идти».
а — список; *б* — двухуровневый морфофонемный префиксный граф.

В отличие от графа, представленного на рис. 6, здесь использована модифицированная структура. Введены виртуальный начальный и концевой узлы, которые необходимы для последующего включения графа в архитектуру автоматического декодера русской речи. Проанализируем сложность двух вариантов представления парадигмы на примере слова «идти». В табл. 2 представлено сравнение списка (рис. 7*а*) и морфофонемного префиксного графа (рис. 7*б*) по числу узлов и ребер. Для одной и той же парадигмы сложность топологии графа уменьшилась почти в два раза с применением ДМПГ.

Таблица 2

Сравнение двух вариантов представления парадигмы

Вариант представления парадигмы лексемы «идти»	**Число узлов**	**Число ребер**
Список цепочек фонем	83	95
Двухуровневый морфофонемный префиксный граф	45	59

4. Представление словаря в виде морфофонемного графа

Используя описанную в предыдущем разделе процедуру преобразования парадигмы слова в компактную форму, получаем набор ДМПГ, описывающий все возможные словоформы словаря. Построение ДМПГ словаря осуществляется путем объединения ДМПГ парадигм слов в единый граф за два этапа. В первую очередь объединяются одинаковые префиксные пути основ, а затем

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

осуществляется слияние на уровне окончаний. Таким образом, будет получен ДМПГ словаря (сеть) со следующими характеристиками:

- Первый уровень представляет собой префиксный граф транскрипций основ, концевыми узлами которого являются графемные представления основ.
- S — число входных узлов, равное числу различных первых фонем в словах из словаря.
- N — число концевых узлов графа первого уровня, равное числу различных транскрипций основ словаря.
- Второй уровень представляет собой префиксный фонемный граф транскрипций окончаний, концевыми узлами которого являются графемные представления окончаний.
- E — число входных узлов второго уровня, равное числу первых фонем в окончаниях.
- K — число концевых узлов графа второго уровня, равное числу различных транскрипций всех окончаний, необходимых для образования всех словоформ словаря.
- Все неконцевые узлы первого и второго уровней графа содержат фонемы.
- Любой путь по двухуровневому графу содержит ровно два «концевых» узла (основу и окончание).
- Число различных путей по графу равно числу всех различных словоформ, которые можно образовать по грамматическим правилам русского языка от имеющегося в словаре списка основ.
- Минимальная и максимальная длины путей по графу равны числу фонем в транскрипциях самой короткой («а!») и самой длинной («высо!капр’ивасхад’и!тел’ствм’и» — 25 фонем) словоформы соответственно.
- В бинарном представлении пути графа, представляющие транскрипции основ, отсортированы по возрастанию («а!»,… «йа!щурками»).

Данный граф может быть применен для распознавания изолированно произнесенных слов. В этом случае последовательность фонем, составляющая некоторое слово w, может быть записана в виде кортежа пройденных узлов по графу:

$$w = \left\langle n_1, n_2, ..., n_i, l_1, n_{i+1}, n_{i+2}, ... n_j, l_2 \right\rangle,$$

где $n_1, n_2, ..., n_i, ..., n_j$ — неконцевые узлы, содержащие фонемы из алфавита. Концевые узлы l_1 и l_2 содержат соответственно некоторую основу и некоторое окончание. Последняя фонема в транскрипции основы находится в узле n_i, а окончания — в узле n_j. Тогда длина основы равна i, а $j - i$ — длина окончания.

Использование строгих грамматических правил при формировании двухуровневого морфофонемного графа обеспечивает построение всех возможных словоформ и исключает возникновение грамматически некорректных комбинаций основы и окончания при декодировании речевого сигнала.

Для использования данного графа в задаче распознавания слитной речи необходимо добавить обратную связь, обеспечивающую генерацию последовательности словоформ с неограниченной длиной. Строго говоря, число слов в последовательности будет зависеть от длины записанного речевого сигнала, и

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

при поступлении последней фонемы гипотеза распознанной фразы (путь по графу) заканчивается последним начатым словом. На рис. 8 показана структурная схема ДМПГ, удовлетворяющая приведенным выше характеристикам и пригодная для декодирования слитной речи.

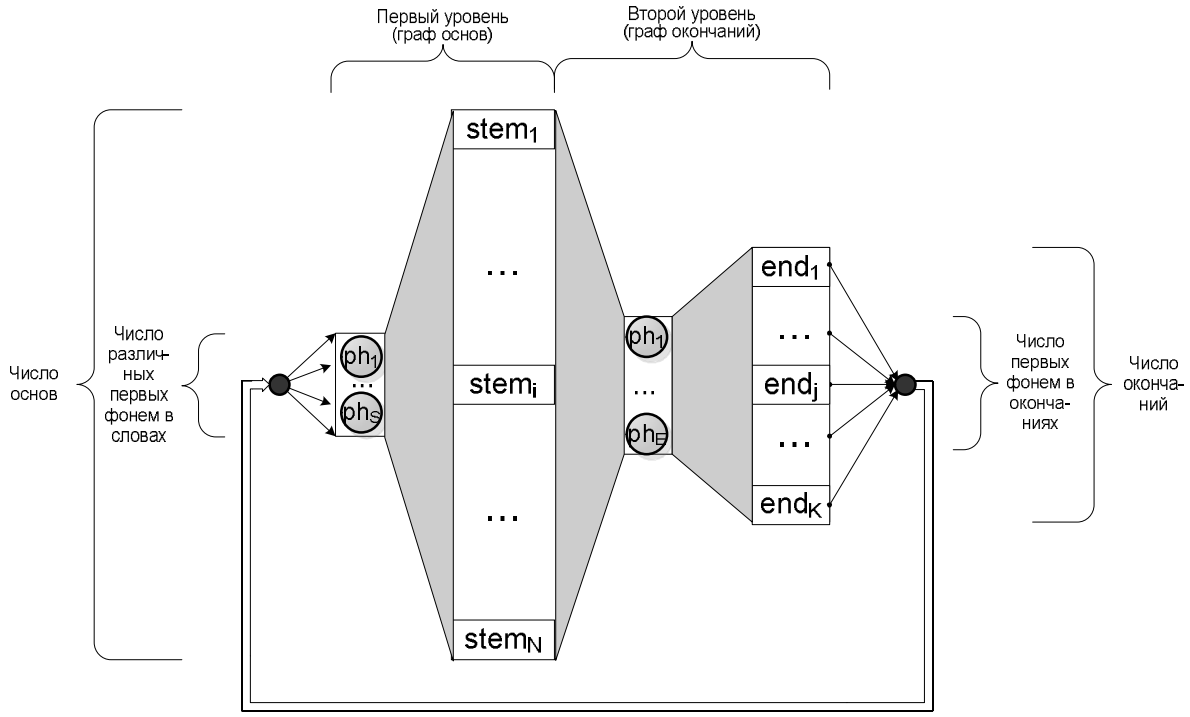


Рис. 8. ДМПГ для декодирования слитной речи.

Как уже было сказано, полученный граф способен генерировать только грамматически правильные слова, т.е. возможные последовательности фонем (произвольные пути по графу) образуют существующие в русском языке словоформы или последовательности словоформ. Однако устная речь существенно отличается от письменной. Более того, лингвисты различают речевую и языковую нормы [12]. Для разговорной речи характерна редукция фонем и целых слов; чаще всего используется некоторый ограниченный словарь, обусловленный конкретной ситуацией; строятся лаконичные логичные высказывания. Также в речи часто употребляются слова, недопустимые в литературном языке, такие как жаргонизмы, вульгаризмы, просторечные слова и т.д. На стиль речи также влияет число участников диалога, условия и цель общения. В целом, можно сказать, что форма и содержание устной речи выбираются по принципу коммуникативной целесообразности, стремясь минимизировать время, затрачиваемое на донесение информации до собеседника.

На данном этапе вопросы семантической связности речи пока не будут затрагиваться, поэтому из предыдущего абзаца сделаем только вывод о том, что изначально на вход декодера поступает цепочка фонем, обозначающая слово, но содержащая ряд ошибок. Ошибки в основном связаны со стилем произношения, акустическими окружающими условиями и особенностями канала передачи данных. Для того чтобы декодер, основанный на ДМПГ, мог обрабатывать любые последовательности фонем и формировать из них грамматически правильные слова, разработанная топология графа закладывается в основу скрытой марковской модели и для всех соседних узлов вводятся вероятности пере-

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

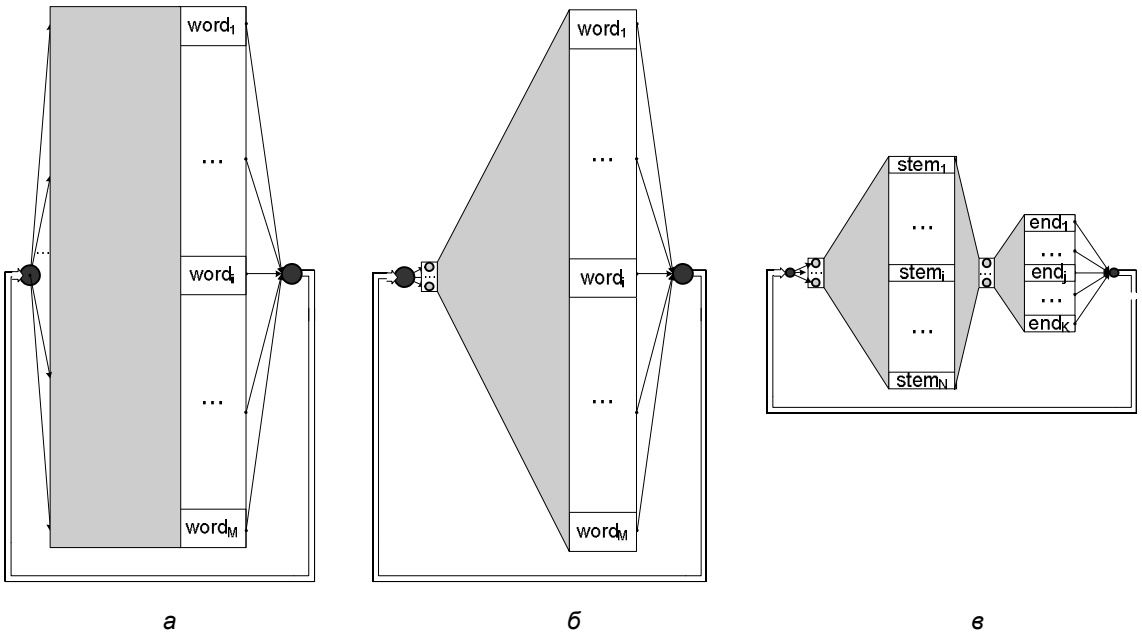
ходов. Полная схема работы декодера построена с учетом хорошо зарекомендовавших себя в области распознавания речи методов: синхронизированного по времени поиска Витерби [6], метода передачи маркеров, методики отсечения маловероятных гипотез, стохастической модели языка, учитывающей комбинации основ [8].

Последующая работа будет посвящена исследованию ассимилятивных процессов, происходящих как внутри слова, так и на стыке двух слов. Особенно подвержены ассимиляции согласные звуки. Для учета этих процессов в граф будут введены дополнительные ребра, описывающие допустимые с точки зрения фонетики транскрипции словоформ. В следующем разделе будут представлены некоторые численные характеристики и преимущества разработанного ДМПГ по отношению к существующим подходам.

5. Сравнение ДМПГ с базовыми моделями

Для оценки предложенного декодера по скорости работы проведем сравнительный анализ ДМПГ с двумя общепринятыми моделями представления словаря: список всех словоформ и лексическое дерево. В конце раздела будет представлено распределение близких по написанию и звучанию слов, которое получается путем анализа длин префиксов в разработанном ДМПГ словаря.

Прежде всего, кратко опишем структуры каждого из трех способов представления словаря. На рис. 9*а* представлен наиболее простой способ хранения словоформ в виде списка, в каждой строке которого содержится слово и его транскрипция. Более компактное представление достигается путем пофонемного объединения идентичных начальных участков слов и построения лексического дерева (рис. 9*б*). Наконец, разработанный способ представления словаря на основе ДМПГ схематично представлен на рис. 9*в.*



а *б* *в*

Рис. 9. Способы представления словаря.
а — список; *б* — лексическое дерево; *в* — ДМПГ.

Из рис. 9 видно, что сложность представления словаря в виде списка пропорциональна числу всех словоформ и средней длине слова. Посредством

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

префиксного дерева достигается значительное сокращение элементов графа. В то же время ДМПГ, построенный по принципам префиксного дерева, сохраняет его преимущества и имеет двухуровневую морфологическую структуру. За счет этого сложность ДМПГ пропорциональна числу основ в словаре.

Для сравнения всех графов использовался один и тот же словарь, параметры которого представлены в табл. 3. Этот словарь содержит список транскрипций словоформ с пометкой границы основы и ссылками на слова.

Таблица 3

Параметры словаря

Параметр словаря	**Число элементов**
Число лексем	69295
Число словоформ (транскрипций)	1521573
Число различных словоформ (транскрипций)	625007
Число различных основ (транскрипций)	113591
Число различных окончаний (транскрипций)	82

Число узлов и дуг, а также плотность графа часто используются для оценки требуемого количества вычислений по графу и во многих случаях являются хорошим показателем сложности топологии. Модели фонем в нашем случае хранятся в узлах, и поэтому плотность графа словаря может быть вычислена как суммарное число узлов, разделенное на число различных словоформ (транскрипций) в словаре. Описанные характеристики графов, построенных по трем разным подходам, представлены в табл. 4. ДМПГ, описывая точно такой же словарь, как и основные модели, почти в 17 раз превосходит их по суммарному числу дуг и узлов, а также имеет в 24 раза меньшую плотность графа.

Таблица 4

Сравнение ДМПГ с двумя основными моделями представления словаря

Критерий сравнения	**Вид представления словаря**		
	Список цепочек фонем	**Лексическое дерево**	**ДМПГ**
Число узлов	6422225	979174	271758
Число дуг	7049324	1604180	528889
Число «листьев»	625007	625007	113673
Суммарное число дуг и узлов	13471549	2583354	800647
Общее сокращение числа дуг и узлов	1	5,21	**16,83**
Плотность графа	10,28	1,57	**0,43**

На данный момент в процессе разработки ДМПГ были сформированы полные парадигмы для номинатива. Статистика по глагольным формам пока не приводилась. Но даже сейчас уже видно, что размерность ДПМГ существенно ниже, чем при других способах представления словаря. После формирования и учета всех форм глаголов (включая множество личных форм, а также причастий и деепричастий) преимущество ДМПГ станет еще более существенным, что показали теоретические расчеты.

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

6. Заключение

Разработка способа компактного представления словаря особенно актуальна для флективных языков с богатой морфологией. Декомпозиция словоформы на основу и окончание по грамматическим правилам посредством разработанного модуля «Диаморф» позволяет хранить словарь в виде префиксного дерева основ и автоматически генерировать произвольную словоформу. При создании модели языка в обучающем тексте выделяются только основы, что позволяет учесть все возможные комбинации словоформ даже при отсутствии таковых при обучении. Предложен декодер слитной русской речи на основе двухуровневого морфофонемного префиксного графа. Процедура его построения сводится к транскрибированию всех словоформ словаря и последующему объединению начальных участков основ и окончаний в двухуровневый ориентированный граф. Прохождение по графу обеспечивает генерацию только грамматически верных словоформ, а использование аппарата скрытых марковских моделей позволяет оценить и выбрать наиболее вероятные гипотезы слов по входной последовательности фонем. Дальнейшая работа будет направлена на расширение графа с учетом ассимилятивных процессов, происходящих внутри и на стыке слов. Это позволит более точно описать возможные способы произношения словоформ и повысить точность распознавания слитной русской речи.

Литература

1. *Зализняк А. А.* Грамматический словарь русского языка. М.: Русские словари, 2003. 800 с.
2. *Kurimo M., Creutz M., Varjokallio M., Arisoy E., Saraclar M.* Unsupervised segmentation of words into morphemes — Morpho challenge 2005 application to automatic speech recognition // Proc. Interspeech 2006. Pittsburgh, USA, 2006. P. 1021–1024.
3. *Kneissler J., Klakow D*. Speech recognition for huge vocabularies by using optimized subword units // Proc. Eurospeech 2001. Aalborg, Denmark, 2001. P. 69–72.
4. *Szarvas M., Furui S.* Finite-state transducer based modeling of morphosyntax with applications to Hungarian LVCSR // Proc. ICASSP'2003. Hong Kong, China, 2003. Vol. 1. P. 368–371.
5. *Карпов А. А., Ронжин А. Л., Ли И. В.* SIRIUS — система дикторонезависимого распознавания слитной русской речи // Известия ТРТУ. 2005. № 10. С. 44–53.
6. *Pražák A., Psutka J., Hoidekr J., Kanis J., Müller L., Psutka, J.* Adaptive language model in automatic online subtitling // Proc. 2nd IASTED International Conference on Computational Intelligence CI 2006. San Francisco, California, USA, 2006. P. 479–483.
7. *Demuynck K., Duchateau J., Van Compernolle D., Wambacq P.* An efficient search space representation for large vocabulary continuous speech recognition // Speech Communication. 2000. Vol. 30, no. 1. P. 37–53.
8. *Carki K., Geutner P., Schultz T.* Turkish LVCSR: Towards better speech recognition for agglutinative languages // Proc. ICASSP-2000. Istanbul, Turkey, 2000. Vol. 3. P. 1563–1566.
9. *Nedevschi S., Patra R., Brewer E.* Hardware speech recognition on low-cost and low-power devices // Proc. of 41st Design Automation Conference. San Diego, USA, 2005. P. 684–689.
10. *Сокирко А. В.* Морфологические модули на сайте www.aot.ru // Диалог-2004. Компьютерная лингвистка и интеллектуальные технологии: Труды междунар. конф. М.: Наука, 2004. 559 с.
11. *Gelbukh A., Sidorov G.* Approach to construction of automatic morphological analysis systems for inflective languages with little effort //. Proc. of CICLing-2003. Lecture Notes in Computer Science. 2003. No. 2588. P. 215–220.
12. *Гойхман О. Я., Надеина Т. М.* Речевая коммуникация. М.: Инфра-М, 2006. 272 с.