УДК 004.522

# АВТОМАТИЧЕСКАЯ ОБРАБОТКА И СТАТИСТИЧЕСКИЙ АНАЛИЗ НОВОСТНОГО ТЕКСТОВОГО КОРПУСА ДЛЯ МОДЕЛИ ЯЗЫКА СИСТЕМЫ РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ

***И. С. Кипяткова,***
*младший научный сотрудник*
***А. А. Карпов,***
*канд. техн. наук, старший научный сотрудник*
*Санкт-Петербургский институт информатики и автоматизации РАН*

*Описывается процесс автоматической обработки текстового корпуса, собранного из новостных лент ряда интернет-сайтов, для создания вероятностной n-граммной модели разговорного русского языка. Приводится статистический анализ данного корпуса, даются результаты по подсчету частоты появления различных n-грамм слов. Представлен обзор существующих типов статистических моделей языка.*

***Ключевые слова*** *— модель языка, текстовый корпус русского языка, автоматическая обработка текста.*

## Введение

Для генерации грамматически правильных и осмысленных гипотез произнесенной фразы распознавателю речи необходима некоторая модель языка или грамматика, описывающая допустимые фразы. Процесс распознавания речи может быть представлен как поиск наиболее вероятной последовательности слов [1]:

$$W = \arg\max_{W} P(W \mid A) = \arg\max_{W} P(A \mid W)P(W),$$

где $P(A|W)$, $P(W)$ — вероятности появления гипотезы по оценке акустической и языковой модели соответственно.

Для многих языков (например, английского) разработаны методы создания моделей языка, которые позволяют повысить точность распознавания речи. Но эти методы не могут быть напрямую применены для русского языка из-за свободного порядка слов в предложениях и наличия большого количества словоформ для каждого слова.

Одной из наиболее эффективных моделей естественного языка является статистическая модель на основе $n$-грамм слов, цель которой состоит в оценке вероятности появления цепочки слов $W = (w_1, w_2, ..., w_m)$ в некотором тексте. $n$-граммы представляют собой последовательность из $n$ элементов (например, слов), а $n$-граммная модель языка используется для предсказания элемента в последовательности, содержащей $n - 1$ предшественников. Эта модель основана на предположении, что вероятность какой-то определенной $n$-граммы, содержащейся в неизвестном тексте, можно оценить, зная, как часто она встречается в некотором обучающем тексте.

Вероятность $P(w_1, w_2, ..., w_m)$ можно представить в виде произведения условных вероятностей входящих в нее $n$-грамм [2]:

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, w_2, \ldots, w_{i-1})$$

или аппроксимируя $P(W)$ при ограниченном контексте длиной $n - 1$:

$$P(w_1, w_2, \ldots, w_m) \cong \prod_{i=1}^{m} P(w_i \mid w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}).$$

Вероятность появления $n$-граммы вычисляется на практике следующим образом:

$$P(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1})},$$

где $C$ — количество появлений последовательности в обучающем корпусе.

Далее описывается процесс сбора и предварительной обработки текста для создания статистической модели русского языка.

### Сбор и автоматическая обработка текстового корпуса

Существуют несколько текстовых корпусов русского языка, например «Национальный корпус русского языка» (www.ruscorpora.ru) и «Корпус русского литературного языка» (www.narusco.ru). Они содержат в основном текстовые материалы конца XX в. различных типов: художественные, публицистические, научные, а также в небольшом объеме стенограммы устной речи. В работе [3] описан новостной корпус, собранный из примерно двух тысяч СМИ-источников объемом 7,3 млрд словоупотреблений. Нами для создания модели языка был собран и обработан новостной текстовый русскоязычный корпус, сформированный из новостных лент последних лет четырех интернет-сайтов: www.ng.ru («Независимая газета»), www.smi.ru («СМИ.ru»), www.lenta.ru («LENTA.ru»), www.gazeta.ru («Газета.ru»). Он содержит тексты, отражающие срез современного состояния русского языка, в том числе разговорного. Пополнение этого корпуса может осуществляться автоматически при обновлении сайтов в режиме он-лайн, что позволяет оперативно добавлять новые появляющиеся в языке слова и переобучать модель языка с учетом новых текстовых данных. Естественный язык, будучи открытой системой, постоянно изменяется с изменением общественной жизни, развитием новых областей знаний, и он-лайн пополнение текстового корпуса позволяет учитывать изменения, происходящие в языке. Общий объем корпуса на данный момент составляет свыше 200 млн словоупотреблений (более 1 ГБ данных).

Автоматическая обработка текстового материала осуществляется следующим образом. Вначале текстовый массив разбивается на предложения, которые должны начинаться либо с заглавной буквы, либо с цифры. При этом учитывается, что в начале предложения могут стоять кавычки. Предложение заканчивается точкой, восклицательным или вопросительным знаком либо многоточием. Кроме того, при разделении текста на предложения учитывается, что внутри предложения могут стоять инициалы и/или фамилии. Формально это похоже на границу раздела двух предложений, поэтому если точка идет после одиночной заглавной буквы, то она не будет считаться концом предложения. Предложения, содержащие прямую и косвенную речь, разделяются на отдельные предложения. При этом возможны три случая:

1) прямая речь идет после косвенной;

2) прямая речь идет до косвенной;

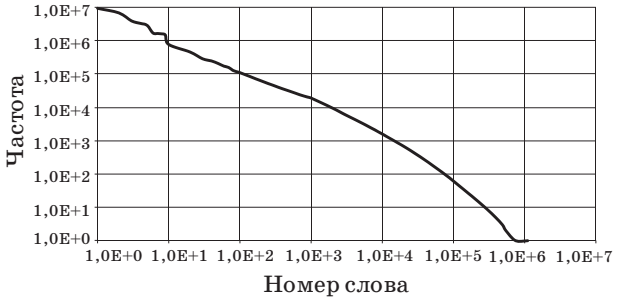
3) косвенная речь находится внутри прямой речи.

В первом случае формальными признаками, при которых происходит выделение прямой и косвенной речи, является наличие двоеточия, после которого следуют кавычки. Во втором случае разделение происходит, если после кавычек стоит запятая, а затем тире. В третьем случае исходное предложение разбивается на три предложения: первое — от кавычек до запятой и тире, второе — то, что находится между первой запятой с тире до второй запятой с тире, третье — от запятой с тире до конца предложения.

После разделения текстового материала на предложения выполняется его нормализация. Происходит удаление текста, написанного в любых скобках, удаление предложений, состоящих из пяти и меньше слов (как правило, это заголовки, составленные не по грамматическим правилам для полных предложений). Затем из текстов удаляются знаки препинания, символы «№» и «#» меняются на слово «номер». Все числа и цифры объединяются в единый класс, который обозначается в результирующем тексте символом «№». За одно число принимается группа цифр, которые могут быть разделены точкой, запятой, пробелом или тире. Также символом «№» обозначаются римские цифры, которые представляют собой совокупность латинских букв *I*, *V*, *X*, *L*, *C*, *D*, *M* и могут быть разделены пробелом или тире. В отдельные классы выделяются интернет-адреса (обозначаются знаком «<>») и адреса E-mail (обозначаются символом «<@>»). В словах, начинающихся с заглавной буквы, происходит замена заглавной буквы на строчную. Если все слово написано заглавными буквами, то замена не делается, так как это слово, вероятно, является аббревиатурой.

### Статистический анализ текстового корпуса

На базе собранного русскоязычного текстового корпуса (более 200 млн словоупотреблений) был создан частотный словарь, размер которого составляет свыше 1 млн уникальных словоформ, а также для данного корпуса определена частота встречаемости различных $n$-грамм слов при $n$ в диапазоне от 2 до 5 лексических элементов. Выполнена проверка соответствия текстового корпуса закону Ципфа (рис. 1). Известно, что закон Ципфа [4] — эмпирическая закономерность распределения частоты слов естественного языка: если все слова языка в достаточно большом осмысленном тексте упорядочить по убыванию частоты их использования, то частота слова в таком спи-

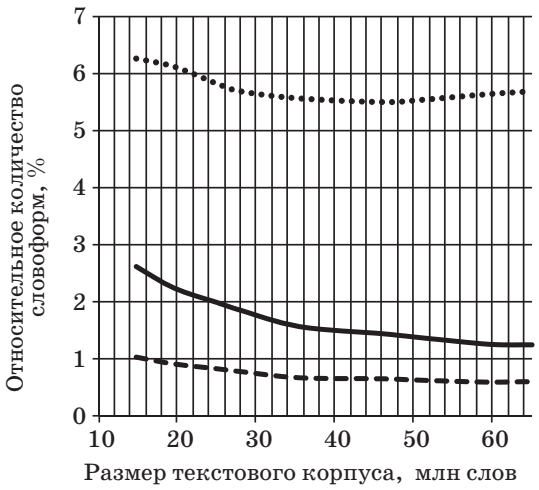



■ ***Рис. 1.*** *Проверка соответствия текстового корпуса закону Ципфа*

ске окажется приблизительно обратно пропорциональной его порядковому номеру. Собранный корпус соответствует закону Ципфа.

Для автоматического распознавания речи необходимо иметь словарь фонематических транскрипций слов. Нами был разработан программный модуль, позволяющий создавать фонематические транскрипции слов автоматически [5]. Для создания транскрипций необходимо наличие базы данных (БД) словоформ русского языка с отметкой ударения. В качестве таковой использовалась БД, созданная путем объединения двух БД, свободно доступных в Интернете: 1) морфологическая БД проекта STARLING (http://starling.rinet.ru); 2) морфологическая БД проекта АОТ (www.aot.ru). Первая БД содержит около 1 млн 800 тыс. различных словоформ, что недостаточно для наших исследований. В этой БД для некоторых сложных слов проставлено также второстепенное ударение. Вторая БД содержит свыше 2 млн 200 тыс. словоформ. Однако здесь, в отличие от первой базы, отсутствует буква *ё* и информация о второстепенном ударении. Поэтому обе БД были объединены, объем получившегося словаря превысил 2 млн 300 тыс. различных словоформ, что является приемлемым для наших задач.

Был проведен анализ того, насколько получившийся объединенный словарь покрывает обрабатываемый текстовый корпус. На рис. 2 представлен график отношения количества уникальных словоформ и словоформ, отсутствующих в фонематическом словаре, к общему количеству словоформ в зависимости от размера текстового корпуса. График показывает, что с ростом размера текстового корпуса относительное количество уникальных словоформ, встречающихся в этом корпусе, падает и составляет 1,2 % при размере текстового корпуса 60 млн словоформ. Для сравнения: относительное количество уникальных словоформ для английского языка при таком же размере текстового корпуса приблизительно равно 0,5 % [6]. Относительное количество уникальных словоформ, отсутствующих в словаре, с размером корпуса практически не изменяется и составляет менее 1 % от общего количества словоупотреблений в тексте. Общее количество внесловарных слов в среднем составляет менее 6 %, — это большое число по сравнению со многими другими языками. В таблице приведено относительное количество внесловарных слов для различных языков [7–9].

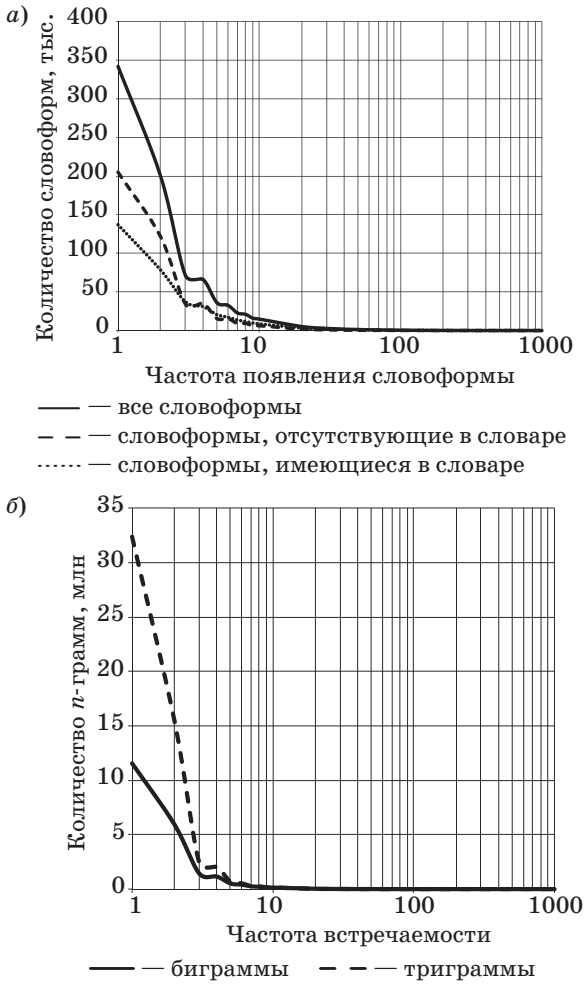



■ ***Рис. 2.*** *Зависимость относительного количества словоформ от размера корпуса*

Графики распределения частот встречаемости униграмм (аналог частотного словаря) (рис. 3, *а*), биграмм и триграмм (рис. 3, *б*) слов показывают, что в текстах присутствует достаточно много редко употребляемых слов. Более 350 тыс. слов встретились только один раз в текстовом корпусе. Кроме того, большая часть словоформ, у которых частота встречаемости меньше 7, отсутствуют в словаре. Как правило, это слова, написанные с опечатками. Поэтому для сокращения списка *n*-грамм и скорости обработки целесообразно

■ *Количество внесловарных слов в текстах для различных языков*

| Язык | Размер словаря, тыс. слов | Количество внесловарных слов, % |
|---|---|---|
| Английский | 300 | 0,2 |
| Чешский | 45 | 5,0 |
| Эстонский | 120 | 6,0 |
| Турецкий | 120 | 5,0 |
| Финский | 400 | 5,0 |
| Литовский | 1000 | 1,9 |
| Русский | 2300 | 6,0 |




■ ***Рис. 3.*** *Распределение частоты встречаемости различных словоформ: а — униграмм; б — биграмм и триграмм*

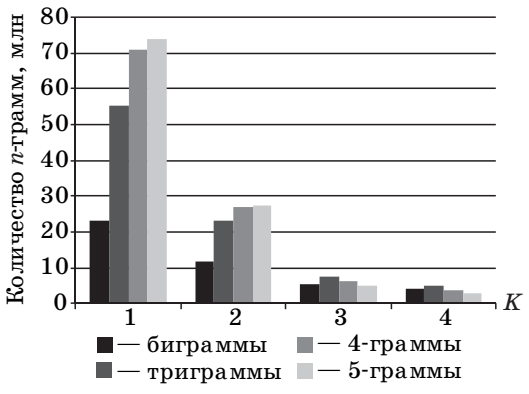
удалять редкие $n$-граммы. Для этого был введен порог $K$; $n$ -граммы, которые встретились меньше $K$ раз, удаляются из списка.

Как уменьшается количество $n$-грамм с ростом $K$, показано на рис. 4. При удалении из списка $n$-грамм, которые встретились только один раз ($K = 2$), список $n$-грамм сократился вдвое, а при $K = 3$ — еще в несколько раз. При дальнейшем увеличении $K$ сокращение было уже незначительным.

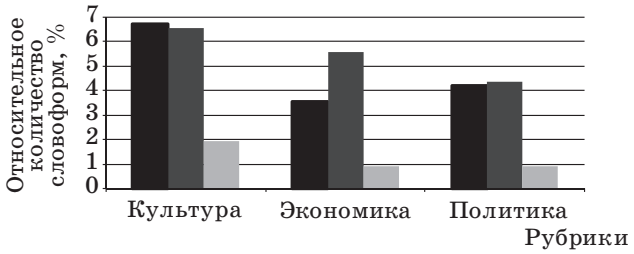
На интернет-сайтах газет новостной материал разделен на различные рубрики. Проанализировано, как изменяется относительное количество уникальных словоформ и словоформ, отсутствующих в словаре, в зависимости от тематики. Наиболее представительными оказались рубрики «Культура», «Экономика» и «Политика», поэтому из всего корпуса были выбраны текстовые данные по 4 млн словоупотреблений для каждой из этих рубрик. На рис. 5 показано распределение относительного количества уникальных словоформ и словоформ, отсутствующих в словаре, по каждой рубрике. Наибольшее количество как уникальных, так и отсутствующих в словаре словоформ было найдено в рубрике «Культура». В рубриках «Экономика» и «Политика» количество уникальных словоформ, отсутствующих в словаре, приблизительно одинаково, однако общее количество отсутствующих в словаре словоформ больше в «Экономике».

Была создана биграммная модель языка с помощью программного модуля обработки и анализа текстов CMU (*Cambridge Statistical Language Modeling Toolkit*) [10]. Поскольку большинство слов с частотой появления меньше 7 отсутствуют в словаре, при создании модели языка был введен порог $K = 7$, т. е. из модели языка удалялись биграммы, у которых значение частоты появления по отношению к размеру корпуса было меньше $3{,}5 \cdot 10^{-8}$. При этом количество уникальных словоформ составило почти 200 тыс., количество биграмм — около 2,1 млн.

Для тестирования созданной модели языка был собран корпус меньшего размера, содержащий текстовый материал новостного сайта www.fontanka.ru («Фонтанка.ru»). На этом тесто-




■ ***Рис. 4.*** *Распределение количества n-грамм при пороге минимальной встречаемости*




— уникальные словоформы
— общее количество словоформ, отсутствующих в словаре
— уникальные словоформы, отсутствующие в словаре

■ ***Рис. 5.*** *Распределение относительного количества уникальных словоформ и словоформ, отсутствующих в словаре, по темам*

вом корпусе была вычислена величина энтропии и коэффициента неопределенности (*perplexity*) статистической модели языка. По определению, информационная энтропия — мера хаотичности информации, неопределенность появления какого-либо символа первичного алфавита. При отсутствии информационных потерь она численно равна количеству информации на символ передаваемого сообщения. Поскольку тексты на естественном языке могут рассматриваться в качестве информационного источника, энтропия вычисляется по следующей формуле [2]:

$$H = -\lim_{m\to\infty}\frac{1}{m}\sum_{w_1,w_2,\ldots,w_m}(P(w_1,w_2,\ldots,w_m)\times \times\log_2 P(w_1,w_2,\ldots,w_m)).$$

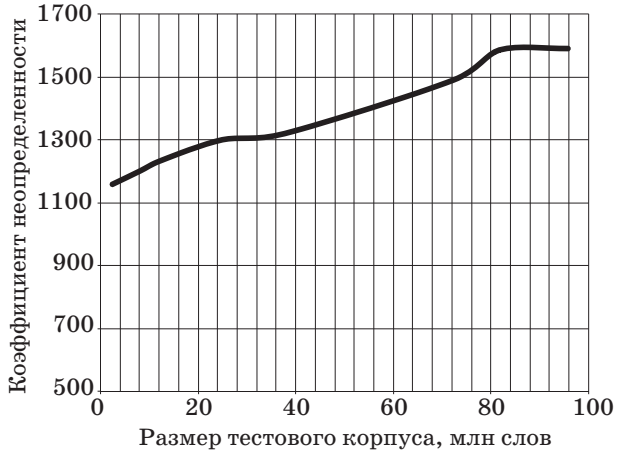
Это суммирование делается по всем возможным последовательностям слов. Но поскольку язык является эргодичным источником информации [2], выражение для вычисления энтропии будет выглядеть следующим образом:

$$\hat{H} = -\frac{1}{m}\log_2 P(w_1,w_2,\ldots,w_m).$$

Коэффициент неопределенности является параметром, по которому оценивается качество $n$-граммных моделей языка, и вычисляется следующим образом [2]:

$$PP = 2^{\hat{H}} = \hat{P}(w_1,w_2,\ldots,w_m)^{-\frac{1}{m}},$$

где $\hat{P}(w_1,w_2,\ldots,w_m)$ — вероятность последовательности слов $w_1$, $w_2$, ..., $w_m$. Коэффициент неопределенности показывает, сколько в среднем различных наиболее вероятных слов может следовать за данным словом. На рис. 6 представлены значения коэффициента неопределенности при различном размере тестового корпуса, величина энтропии составляет 1,18–1,64 бит/слово, относительное количество новых слов в этом корпусе находится в пределах от 1,1 до 1,7 % при размерах корпуса от 2,5 до 95,8 млн словоформ. Полученные значения являются достаточно большими. Например, для английского языка при размере словаря в 200 тыс. слов коэффициент неопределенности равен 232 [6], при этом энтропия будет приблизительно равна 7,9 бит/слово, а относительное количество новых слов составляет 0,31 % для тестового корпуса объемом 1,12 млн слов.




■ ***Рис. 6.*** *Зависимость значения коэффициента неопределенности от размера тестового корпуса*

## Разновидности статистических моделей языка

В данном разделе рассматриваются возможные варианты построения моделей языка, основанных на статистическом анализе текста.

Модели, основанные на классах (*class-based models*), используют функцию, которая отображает каждое слово $w_i$ на класс $c_i$: $f$: $w_i \to f(w_i) = c_i$. В этом случае оценка условной вероятности может быть аппроксимирована по $n$-грамме класса [9]:

$$P(w_i \mid w_{i-n+1},\ldots,w_{i-1}) = P(w_i \mid c_i)P(c_i \mid c_{i-1+1},\ldots,c_{i-1}).$$

Функция отображения слова на класс может быть определена вручную с использованием некоторой морфологической информации (например, информации о части речи). Также существуют методы, которые помогают определить функцию отображения автоматически по текстовому корпусу.

Интервальные модели языка (*distance models*) помогают включить больший контекст, чем $n$-граммы, но величина коэффициента неопределенности модели остается того же порядка, как у $n$-грамм. Например, биграммная интервальная модель может быть задана следующим образом [9]:

$$P(w_i \mid w_{i-M+1},\ldots,w_{i-1}) = \sum_{m=1}^{M-1}\lambda_m P_m(w_i \mid w_{i-m}),$$

где $M$ — предопределенное число моделей; $\lambda_m$ — весовые параметры модели при условии $\sum_{m=1}^{M-1}\lambda_m = 1$; $P_m(w_i|w_{i-m})$ — биграммная модель с пропуском $m-1$. Значение весовых коэффициентов $\lambda_m$ определяется как зависимость от расстояния от слова $w_i$ до слова $w_{i-m}$ (с увеличением расстояния до слова величина весового коэффициента уменьшается).

Триггерные модели (*trigger models*) — это другой тип моделей, которые моделируют взаимоотношение пар слов в более длинном контексте. В этом методе появление инициирующего слова в истории увеличивает вероятность другого слова, называемого целевым, с которым оно связано.

Вероятность пар слов может быть определена следующим образом [9]:

$$P_{a\to b}(b \mid a \in h) = \frac{C(a \in h, b)}{C(a \in h)},$$

где $a$ — инициирующее слово; $b$ — целевое слово; $h$ — история некоторого ограниченного размера для слова $b$, т. е. слова, предшествующего в тексте слову $b$; функция $C$ определяет подсчет события в текстовом корпусе.

Упрощенной версией триггерных пар является кэш-модель (*cache model*). Кэш-модель увеличивает вероятность появления слова в соответствии с тем, как часто данное слово употреблялось в истории, поскольку считается, что, употребив конкретное слово, диктор будет использовать это слово еще раз либо из-за того, что оно является характерным для конкретной темы, либо потому, что диктор имеет тенденцию использовать это слово в своем лексиконе. Обычная униграммная кэш-модель может определяться как [9]

$$P_C(w_i \mid h) = \frac{C(w_i, h)}{C(h)} = \frac{\sum_{j=i-D}^{i-1} I(w_i = w_j)}{\sum_{j=i-D}^{i-1} I(w_j \in V)},$$

где $D$ — размер истории $h$; $I$ — индикаторная функция; $V$ — словарь модели языка.

Другим типом модели языка является модель на основе набора тем (*topic mixture models*). Текстовый корпус вручную или автоматически делится на предопределенное число тем, и языковые модели создаются отдельно для каждой темы. Полная модель может определяться как [9]

$$P_{TM}(w_i \mid h_i) = \sum_{j=1}^{M} \lambda_j P_j(w_i \mid h_i),$$

где $M$ — число тем; $P_j$ — модель темы $j$ с весом модели $\lambda_j$.

Модели, основанные на частях слов (*particle-based models*), используются для языков с богатой морфологией, например флективных языков [9]. В этом случае слово $w$ разделяется на некоторое число $L(w)$ частей (морфем) с помощью функции $U$: $w \to U(w) = u^1, u^2, ..., u^{L(w)}$, $u^i \in \Psi$, где $\Psi$ — набор частей слова. Разделение слов на морфемы можно производить двумя путями: при помощи словарных и алгоритмических методов [11]. Преимуществом алгоритмических методов является то, что они опираются лишь на анализ текста и не используют никаких дополнительных знаний, что позволяет анализировать текст на любом языке. Преимущество словарных методов заключается в том, что они позволяют получить правильное разбиение слов на морфемы, а не на псевдоморфемные единицы (как в алгоритмических методах), что может быть использовано далее на уровне постобработки гипотез распознавания фраз.

Хотя, по определению, $n$-граммные модели языка хранят только $n$ слов, существуют модели, которые не ограничивают последовательности слов до определенного $n$, а вместо этого хранят различные последовательности разной длины. Такие модели называют $n$-граммами переменной длины (*varigrams*) [2]. По существу они могут рассматриваться как $n$-граммные модели с большим $n$ и такими принципами сокращения длины моделей, которые сохраняют только небольшой поднабор всех длинных последовательностей, встретившихся в обучающем тексте.

Автор работы [12] предлагает дальнодействующую триграммную модель, которая представляет собой триграммную модель с разрешенными связями между словами, находящимися не только в пределах двух предыдущих слов, но и на большем расстоянии от предсказываемого слова. Лежащая в основе «грамматика» представляет собой множество пар слов, которые могут быть связаны вместе через несколько разделяющих слов.

В статье [13] предлагаются составные языковые модели. Автор вводит понятие категорной языковой модели и, в частности, категорных $n$-грамм. Каждому слову в словаре приписываются 15 атрибутов, определяющих грамматические свойства словоформы. Множество значений атрибутов определяет класс словоформы. Каждое слово в предложении рассматривается как его начальная форма и морфологический класс. В итоге грамматика разбивается на две составляющие: изменяемую часть (основанную на морфологии) и постоянную часть (основанную на начальных формах слов), которая строится как $n$-граммная языковая модель.

Для решения проблемы многозначности слов при автоматическом переводе с русского языка на латышский [14] вместо биграмм используются синтаксические отношения и связи между парами элементов предложения. Из корпуса текстов латышского языка с помощью парсера выбираются синтаксически связанные пары слов. Определяется частота каждой уникальной пары, после чего вычисляется вероятность появления данной синтаксической пары.

## Заключение

Текстовый материал для статистической обработки был взят из интернет-сайтов четырех электронных газет. Таким образом, корпус, предназначенный для создания модели языка, основывается на текстах с большим количеством стенограмм выступлений и прямой речи, отражаю-

щих особенности современного языка, а не на литературных текстах, которые крайне далеки от разговорной речи. Разработанная методика сбора текстового материала позволяет при обновлении интернет-сайтов оперативно дополнять текстовый корпус и затем переобучать модель языка в режиме он-лайн, учитывая тем самым изменения, происходящие как в самом языке, так и в контексте текущих событий. Однако использование интернет-материалов имеет и ряд недостатков, главным из которых является наличие в текстах опечаток. Кроме того, в таких текстах присутствует много имен собственных, большинство из которых в разговорной речи встречается редко. Из-за этого возрастает объем созданных в результате обработки текста *n*-грамм.

Статистические данные, полученные при обработке текста, будут в дальнейшем использованы для создания модели русского языка для системы распознавания речи. Проведенный анализ показывает, что большинство стандартных методов создания моделей языка не подходят для русского языка. В русском языке очень велико соотношение уникальных слов к размеру текстового корпуса. Для решения данной проблемы целесообразно создавать модель языка, основываясь на начальных формах слов или используя основы слов. Это позволит сократить размер словаря распознавателя и списков *n*-грамм.

Данное исследование поддержано Советом по грантом Президента РФ (проект МК-64898.2010.8), Минобрнауки РФ в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» (госконтракт № П2579), фондом РФФИ (проекты № 08-08-00128 и 09-07-91220-СТ), а также фондом «Научный потенциал» (договор № 201).

## Литература

1. **Rabiner L., Juang B.-H.** Fundamentals of Speech Recognition. — Prentice Hall, 1995. — 507 p.
2. **Moore G. L.** Adaptive Statistical Class-based Language Modelling: PhD thesis. — Cambridge University, 2001. — 193 p.
3. **Баглей С. Г., Антонов А. В., Мешков В. С., Суханов А. В.** Статистические распределения слов в русскоязычной текстовой коллекции: Материалы Междунар. конф. «Диалог 2009». М., 2009. С. 13–18.
4. **Gelbukh A., Sidorov G.** Zipf and Heaps Laws' Coefficients Depend on Language: Proc. Conf. on Intelligent Text Processing and Computational Linguistics, 2001, Mexico City//Lecture Notes in Computer Science. Springer-Verlag, 2001. № 2004. P. 332–335.
5. **Кипяткова И. С., Карпов А. А.** Разработка и оценивание модуля транскрибирования для распознавания и синтеза русской речи // Искусственный интеллект. 2009. № 3. С. 178–185.
6. **Whittaker E. W. D.** Statistical Language Modelling for Automatic Speech Recognition of Russian and English: PhD thesis. — Cambridge University, 2000. — 140 p.
7. **Ircing P., Hoidekr J., Psutka J.** Exploiting Linguistic Knowledge in Language Modeling of Czech Spontaneous Speech: Proc. of LREC 2006. Paris: ELRA, 2006. P. 2600–2603.
8. **Kurimo M.** et al. Unlimited vocabulary speech recognition for agglutinative languages: Proc. of the Human Language Technology Conf. of the North American Chapter of the ACL. N. Y., 2006. P. 487–494.
9. **Vaičiūnas A.** Statistical Language Models of Lithuanian and Their Application to Very Large Vocabulary Speech Recognition. Summary of Doctoral Diss. / Vytautas Magnus University. — Kaunas, 2006. — 35 p.
10. **Clarkson P., Rosenfeld R.** Statistical language modeling using the CMU-Cambridge toolkit: Proc. EUROSPEECH. Rhodes, Greece, 1997. P. 2707–2710.
11. **Kurimo M.** et al. Unsupervised decomposition of words for speech recognition and retrieval // Speech and Computer: Proc. of 13th Intern. Conf. SPECOM'2009. St. Petersburg, 2009. P. 23–28.
12. **Протасов С. В.** Вывод и оценка параметров дальнодействующей триграммной модели языка: Материалы Междунар. конф. «Диалог 2008». М., 2008. С. 443–449.
13. **Холоденко А. Б.** О построении статистических языковых моделей для систем распознавания русской речи // Интеллектуальные системы. 2002. Т. 6. Вып. 1–4. С. 381–394.
14. **Горностай Т., Васильев А., Скадиньш Р., Скадиня И.** Опыт латышско-русского машинного перевода: Материалы Междунар. конф. «Диалог 2007». М., 2007. С. 137–146.