УДК 004.934.5

Л.И. Цирульник, Д.А. Покладок

ГРАММАТИЧЕСКИЙ СЛОВАРЬ И ПРАВИЛА ОПРЕДЕЛЕНИЯ СЛОВЕСНОГО УДАРЕНИЯ ДЛЯ СИНТЕЗА РЕЧИ ПО ТЕКСТУ НА МОБИЛЬНЫХ УСТРОЙСТВАХ

Описываются исследования, направленные на выявление закономерностей расстановки ударений в словах русского языка. Предлагается алгоритм расстановки словесных ударений в произвольных текстах. Тестирование полученных результатов показывает, что с использованием разработанных правил точность расстановки словесных ударений уменьшается всего на 1,4 % при уменьшении объема грамматического словаря в восемь раз. Такое уменьшение объема позволяет использовать словарь в системе синтеза речи для мобильных устройств, характеризующихся малым объемом памяти.

Введение

Системы синтеза речи по тексту к настоящему моменту достигли высокого качества как по критериям разборчивости и естественности синтезируемого голоса, так и по техническим характеристикам, что способствует их широкому применению в центрах обработки вызовов, при управлении сложными объектами, для создания аудиокниг и т. д. Широкое распространение может получить использование систем синтеза речи на мобильных устройствах, таких как карманные персональные компьютеры, смартфоны, мобильные телефоны. Примерами приложений могут служить озвучивание SMS-сообщений [1] и указаний автомобильной навигационной системы [2], чтение писем электронной почты [3].

Большинство современных систем синтеза речи для мобильных устройств, например Acapela TTS for Windows Mobile [4], Nuance TALKS [5], Mobile Speak [6] и др., работают на смартфонах под управлением операционных систем Windows Mobile или Symbian. Однако смартфоны, на которых возможна работа этих систем синтеза речи, составляют, по различным подсчетам, только 10–15 % рынка мобильных телефонов [7, 8].

Основная масса мобильных аппаратов, используемых в настоящее время, характеризуется низким быстродействием и относительно небольшим объемом памяти. В то же время современные системы синтеза речи требуют большого объема памяти для хранения лингвистических и акустических ресурсов, что не позволяет напрямую «перенести» существующие системы на мобильные платформы. При этом одним из основных информационных ресурсов системы синтеза речи является грамматический словарь, который используется для определения словесного ударения и лексико-грамматических характеристик слов входного текста. Объем электронного грамматического словаря достигает 50 МБ [9].

Целью исследований, описанных в настоящей статье, является выявление закономерностей расстановки словесного ударения с целью уменьшения объема словаря. Подобные исследования проводились и ранее. Например, в работе [10] приводятся эвристические правила определения словесного ударения, основанные на морфологических признаках слов. Однако приводимые правила выполняются далеко не всегда, их использование требует корректировки и уточнения. В работе [11] описан вероятностный метод определения позиции словесного ударения, учитывающий стандартные префиксы и словоизменительные окончания русского языка. Этот метод требует нахождения возможных грамматических характеристик и части речи слова для определения в нем позиции ударения. Он показывает более точные результаты, чем предыдущий, однако не может быть применен на мобильных устройствах, поскольку необходимость нахождения морфологических и грамматических характеристик слов потребует либо большого объема памяти, либо больших вычислительных ресурсов, что практически нивелирует преимущества автоматической расстановки позиции словесного ударения. В настоящей статье описан метод определения словесного ударения, основанный на эвристических и статистических правилах.

1. Расстановка ударений в словаре

Исследования проводились в два этапа. На первом этапе проверялись и корректировались эвристические правила расстановки словесных ударений, на втором – выявлялись статистические правила расстановки ударений.

Набор эвристически выявленных закономерностей включает 47 правил (содержащих 171 подправило), которые формулируются, например, так: «*Если слово заканчивается на «-льон-» (плюс окончание), то ударение падает на этот слог»* или «*В словах, содержащих суффиксы «-оват-», «-еват-», ударение падает на вторую гласную суффикса»*.

В качестве материала для исследований был взят грамматический словарь А.А. Зализняка [12]. Словарь содержит около 100 000 слов и отражает (с помощью специальной системы условных обозначений) современное словоизменение, т. е. склонение существительных, прилагательных, местоимений, числительных и спряжение глаголов. Для исследования была использована электронная версия словаря [9], в котором для каждого слова указана позиция ударения. Словарь был сформирован путем построения парадигм всех слов и последующего удаления дубликатов. С целью исключения неоднозначной интерпретации из словаря были удалены омографы, а также клитики и частично-ударные слова. Объем полученного словаря составил 2 070 816 уникальных слов.

Эвристические правила были запрограммированы, затем для каждого правила было выявлено количество слов в словаре, удовлетворяющих ему, а также количество слов, для которых данное правило не выполняется. Если правило выполнялось менее чем в 80 % случаев, оно исключалось из дальнейшего рассмотрения. В результате было оставлено 41 правило, содержащее 125 подправил.

Далее были выявлены и удалены правила, полностью и почти полностью поглощаемые другими. Например, исследования показали, что правилу «*Если слово заканчивается на «-ованн-» плюс окончание, то ударение падает на первую гласную подстроки»* удовлетворяют 10 812 слов, из которых 10 786 удовлетворяют правилу «*Если слово заканчивается на подстроку «-иров-» плюс окончание и перед подстрокой в слове стоит по меньшей мере один слог, то ударение падает на первую гласную подстроки»*. Второму из правил удовлетворяет 46 451 слово. Таким образом, первое правило почти полностью поглощается вторым, но не наоборот. В результате удаления поглощаемых правил осталось 36 правил, содержащих 98 подправил.

В ходе исследований было выявлено, что некоторые слова удовлетворяют условиям двух правил, но при этом позиция ударения в них соответствует одному из правил, но не соответствует другому. Например, слово «*биологи»* удовлетворяет условиям следующих двух правил, но только первое из них выполняется: «*Если слово заканчивается на подстроку «-олог» плюс окончание, то ударение падает на первую гласную подстроки»* и «*В словах, заканчивающихся на подстроку «-логи-» плюс окончание, ударение падает на первую гласную подстроки»*. Для того чтобы результирующим правилам удовлетворяло максимально возможное число слов, для каждого правила был определен приоритет. При этом из двух правил менее приоритетным считалось то, большее количество исключений из которого удовлетворяет другому правилу. Если же исключения из правила не удовлетворяют ни одному другому правилу, то оно имеет наивысший приоритет.

После такого рода уточнения и корректировки эвристических правил была вычислена статистика выполнения этих правил на материале словаря. В итоге было установлено, что из 2 070 816 слов эвристическим правилам удовлетворяют 786 179 слов, или 38 % объема словаря. Данные слова были удалены из словаря, в результате чего его объем составил 1 284 637 слов.

На следующем этапе исследований выявлялись, как уже было сказано выше, статистические характеристики позиций ударения в словах. Для этого все слова были разбиты на классы в зависимости от количества слогов в них, т. е. образовался класс двусложных, трехсложных и т. д. слов (максимальное количество слогов в словах равнялось 11). Далее для каждого класса вычислялось количество слов, в которых ударение падает на первый слог, второй слог и т. д. вплоть до числа слогов в данном классе. Результаты вычислений представлены в таблице, где каждая строка характеризует определенный класс (т. е. слова с определенным количеством слогов), каждый столбец – количество слов из этого класса, в которых ударение падает на первый,

второй и т. д. слог. Например, общее количество трехсложных слов в словаре равно 235 525. Из них, как видно из третьей строки таблицы, количество слов с ударением на первый слог – 58 371, с ударением на второй слог – 126 305 и с ударением на третий слог – 50 849.

Статистика распределения позиций ударения в зависимости от количества слогов в слове

Кол-во слогов слова	Количество слов, в которых ударение падает на слог с номером										
	1	2	3	4	5	6	7	8	9	10	11
2	34272	25158	–	–	–	–	–	–	–	–	–
3	58371	126305	50849	–	–	–	–	–	–	–	–
4	46585	141078	165881	14241	–	–	–	–	–	–	–
5	22939	80725	156218	44445	2675	–	–	–	–	–	–
6	4576	22363	86140	44435	7541	422	–	–	–	–	–
7	415	2062	18383	22862	8121	1140	34	–	–	–	–
8	20	71	896	4109	4441	1165	111	0	–	–	–
9	8	0	0	171	923	613	110	1	0	–	–
10	0	0	0	15	17	91	52	5	0	0	–
11	0	0	0	0	0	0	0	1	0	0	0

На основании полученных результатов было сформулировано следующее правило: ударным слогом в слове из *n* слогов принимается тот, на который чаще всего падало ударение в исследуемых словах. Например, в трехсложных словах, согласно этому правилу, ударным принимается второй слог, в четырехсложных словах – третий слог и т. д. Можно утверждать, что в общем случае ударение падает на серединный или следующий за ним слог слова.

Все слова, ударения в которых соответствуют сформулированному правилу, были удалены из словаря. Общее количество удаленных слов равнялось 531 867, или 41,4 %.

В результате выявления и использования эвристических и статистических правил объем словаря был уменьшен на 64 % и составил 752 770 слов.

2. Расстановка ударений и частоты встречаемости слов в корпусе текстов

Результаты, полученные на основе словаря, могут оказаться не совсем объективными, поскольку не все слова, встречающиеся в текстах на естественном языке, присутствуют в словаре. Для получения более объективных результатов, а также для проверки точности эвристических правил необходимо провести исследования на корпусе текстов, причем в корпусе должны быть указаны словесные ударения.

Кроме того, корпус текстов можно использовать для дальнейшего уменьшения содержимого словаря, а именно для удаления из него редко встречающихся (или вообще неупотребимых) слов. К таким словам относятся, в частности, словоформы, образование которых затруднительно (некоторые краткие прилагательные, например «деревенск», «геройск», «величайш»; некоторые существительные в родительном падеже множественного числа, например «мечт», «брюзг», «башок»; некоторые страдательные причастия и т. п.), а также устаревшие слова (например «уповать», «дабы» и т. п.). Корпус текстов, используемый для удаления из словаря неупотребимых слов, должен содержать максимальное количество словоформ, но не обязательно должен иметь, в отличие от первого корпуса, указатели словесных ударений.

Для решения второй задачи использовался Национальный корпус русского языка (НКРЯ) [13], который характеризуется представительностью и содержит художественные, публицистические, учебные, научные, деловые, разговорные, диалектные и другие тексты общим объемом более 190 млн словоупотреблений. Для решения первой задачи использовался подкорпус НКРЯ со снятой омонимией (и с указанными позициями словесных ударений) объемом около 6 млн словоупотреблений.

Проверка точности эвристических правил осуществлялась путем выполнения запросов к подкорпусу НКРЯ и обработки полученных наборов слов. Результат проверки представлен на

рис. 1, где по оси абсцисс показаны номера правил и подправил, расположенные по приоритету, по оси ординат – количество слов, позиция ударения в которых соответствует данному правилу.

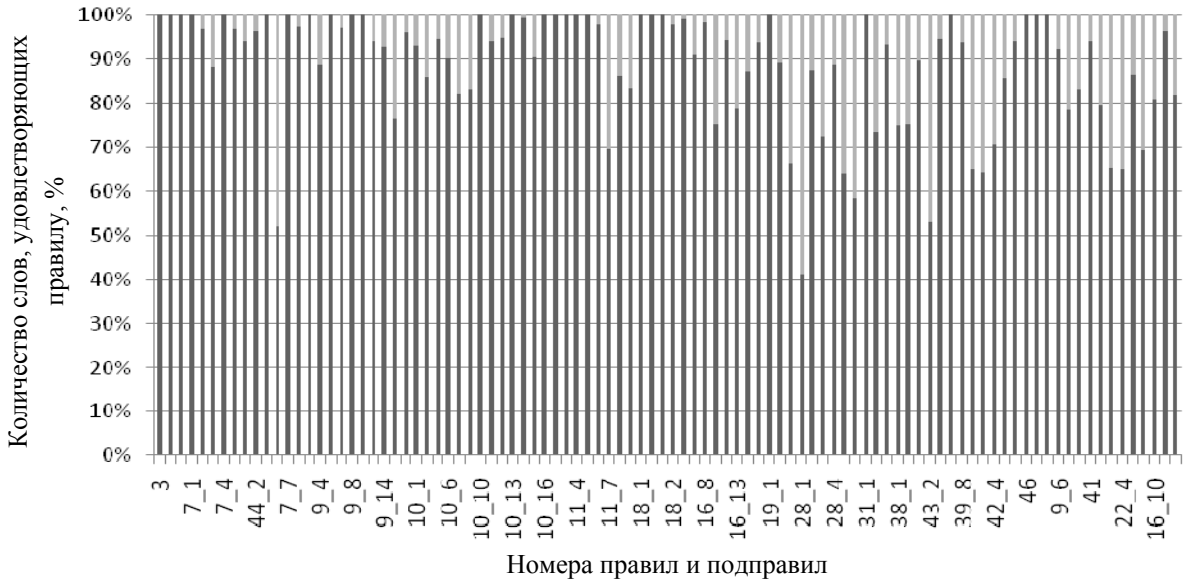

Рис. 1. Статистические характеристики выполнения эвристических правил в корпусе текстов

Как видно из рис. 1, для 20 правил из 98 количество слов, удовлетворяющих этим правилам, не превышает 80 %. Условия данных правил были уточнены, после чего были заново вычислены статистические показатели выполнения всех правил. Для трех правил процент выполнения не превысил 80, и эти правила были удалены.

При проверке точности эвристических правил в НКРЯ было найдено 5 295 слов (более 11 300 словоупотреблений с учетом частоты встречаемости слов), отсутствующих в словаре. Из них 940 слов (2 193 словоупотребления) явились исключениями из правил, и в них были неправильно определены позиции ударений. Однако в процентном соотношении количество ошибок составило 1,8 % от всех проверяемых слов (или 1 % с учетом частоты встречаемости этих слов). Таким образом, точность эвристических правил с учетом частоты встречаемости слов достигает 99 %.

Для определения неупотребимых слов была вычислена частота встречаемости слов словаря в НКРЯ (рис. 2).

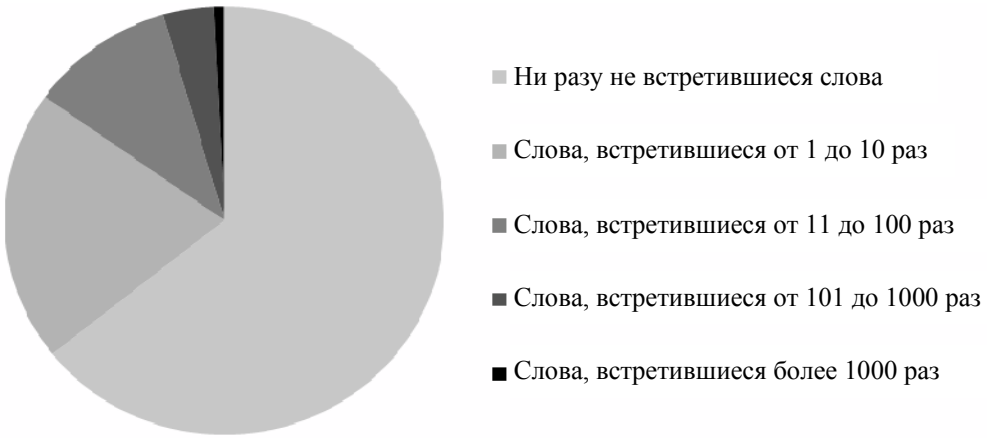

Рис. 2. Результаты вычисления частоты встречаемости слов словаря в НКРЯ

Вычисления показали, что из 752 770 слов словаря 486 627 слов ни разу не встретились в НКРЯ. Следовательно, они могут быть удалены из словаря. Результирующий объем словаря составил 266 143 слова, или 12 % от исходного объема.

3. Алгоритм определения словесного ударения

На основании результатов исследований был разработан алгоритм определения позиции ударения в произвольном слове. Он состоит из трех этапов: поиск по словарю, поиск по эвристическим правилам, поиск по статистическим правилам. Если на каком-то этапе позиция ударения определена, поиск прекращается.

Входные данные алгоритма: произвольное слово, в котором необходимо определить позицию ударения.

Ресурсы алгоритма:

– грамматический словарь с указанием словесных ударений;

– набор эвристических и статистических правил определения позиции ударения.

Выходные данные: слово с указанной позицией ударения.

Структура алгоритма:

Шаг 1. Осуществляется поиск входного слова в словаре. Если слово найдено, в нем устанавливается ударение согласно отмеченному в словаре и происходит переход к шагу 5.

Шаг 2. Для каждого i от 1 до n , где n – количество эвристических правил,

{

если слово удовлетворяет условию i-го правила,

то ударение устанавливается согласно i-му правилу и

происходит переход к шагу 5.

}

Шаг 3. Вычисляется количество слогов в слове j.

Шаг 4. Ударение устанавливается согласно статистическому правилу для слов, имеющих j слогов.

Шаг 5. Конец алгоритма.

4. Тестирование точности алгоритма определения позиции ударения

Тестирование проводилось на части НКРЯ, предоставляемой для свободного использования. Эта часть является случайной выборкой предложений из корпуса со снятой омонимией (с указанными позициями ударений) объемом 180 тысяч словоупотреблений.

Из текстов подкорпуса были предварительно удалены омографы, цифры, символы, не являющиеся буквами русского алфавита, и ошибочные слова (без указанных позиций ударений).

Тестировалась корректность алгоритма определения позиций ударения с использованием исходного словаря (объемом более 2 млн словоформ) и сокращенного словаря (объемом немногим более 266 тыс. словоформ).

Результаты тестирования для слов и словоупотреблений в подкорпусе представлены на рис. 2 и 3 соответственно. По оси абсцисс на рисунках цифрами 1, 2, 3 показаны этапы выполнения алгоритма: 1 – поиск в словаре, 2 – определение позиции ударения в соответствии с эвристическими правилами, 3 – определение позиции ударения в соответствии со статистическими правилами. Для каждого этапа показано процентное соотношение слов (на рис. 2) или словоупотреблений (на рис. 3), для которых на этом этапе была правильно либо ошибочно определена позиция ударения. Доли правильно определенных ударений показаны светло-серым цветом, доли ошибочно определенных – черным. Например, на рис. 2, *а* показано, что на этапе 1 (поиск в словаре) были правильно определены позиции ударения для 93,4 % слов подкорпуса. На этапе 2 (применение эвристических правил) были правильно определены ударения в 1,8 % слов подкорпуса и ошибочно – в 0,2 % слов. На этапе 3 (применение статистических правил) из оставшихся 4,6 % слов для половины (2,3 %) позиции ударения были определены правильно и для половины – ошибочно.

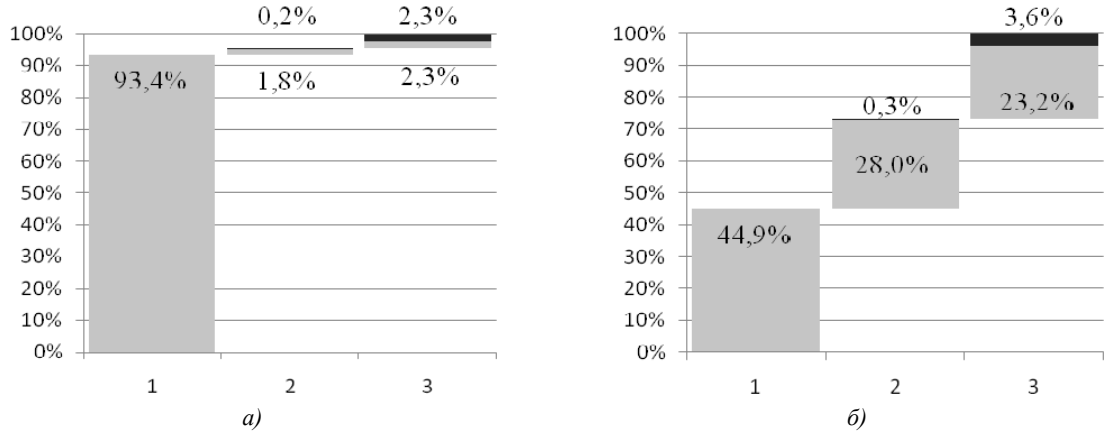

Рис. 2. Точность определения позиции ударения для слов подкорпуса с использованием: *а)* полного словаря; *б)* сокращенного

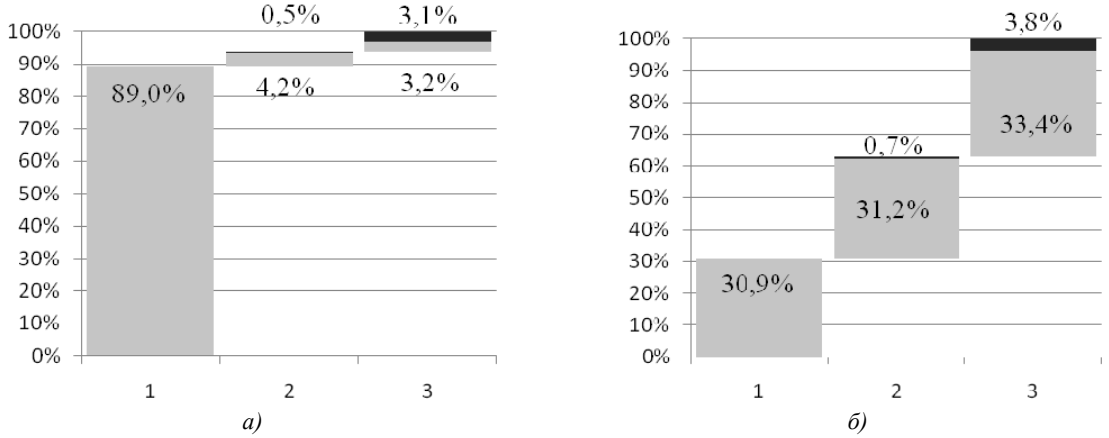

Рис. 3. Точность определения позиции ударения для словоупотреблений подкорпуса с использованием: *а)* полного словаря; *б)* сокращенного

Из рисунков видно, что в полном словаре можно найти более чем в два раза больше слов (и почти в три раза больше словоупотреблений), чем в сокращенном. Эвристические правила могут быть применены к 2 % слов (или 4,7 % словоупотреблений), оставшихся после поиска в полном словаре, в то время как среди слов, оставшихся после поиска в сокращенном словаре, 28,3 % (31,9 % словоупотреблений) удовлетворяют условиям эвристических правил. Процент правильно определенных ударений с использованием эвристических правил довольно высок как в случае использования полного, так и в случае использования сокращенного словарей. Рисунки также показывают, что применение статистических правил дает удовлетворительные результаты: в 23,2 % слов (33,4 % словоупотреблений) ударения определены правильно и только в 3,6 % слов (3,8 % словоупотреблений) – ошибочно при использовании на предыдущих этапах сокращенного словаря.

Рисунки также наглядно демонстрируют, что точность алгоритма определения позиции ударения с использованием сокращенного словаря составляет 95,4 % (или, с учетом частоты встречаемости слов, 96,1 %), что меньше точности того же алгоритма с использованием полного словаря всего на 0,9 % (или, с учетом частоты встречаемости слов, на 1,4 %).

5. Особенности структуры электронного словаря и программной реализации процедуры поиска по словарю

При разработке формата хранения словаря ставились две задачи: уменьшение объема словаря и ускорение процедуры поиска по словарю. Поиск осуществляется с помощью хэш-таблиц, что обеспечивает вычислительную сложность процедуры поиска порядка O(1) [14].

Для уменьшения объема электронного словаря сохранялись не сами слова, а соответствующие им хэш-коды. Для каждого слова словаря был сгенерирован 26-битовый хэш-код и 6-битовый код варианта позиции ударения. В результате под каждое слово было выделено 4 байта. Данный список был отсортирован в порядке убывания и разделен на 256 файлов. Именами этих файлов являются старшие байты хэш-кодов слов в шестнадцатеричном представлении. Таким образом, хранение старшего байта осуществляется в имени файла, а остальные 3 байта расположены в одноименном файле в порядке убывания для последующего быстрого поиска.

В результате программная версия словаря, содержащего 266 143 слова, составила 789 кБ.

Заключение

В результате исследований были уточнены эвристические правила определения позиции ударения в словах русского языка, а также сформулированы статистические правила. Использование этих правил позволило сократить объем грамматического словаря с более 2 млн слов до чуть более 260 тыс. слов, или в восемь раз.

Предложен алгоритм определения словесного ударения, использующий эвристические и статистические правила расстановки ударений, а также грамматический словарь (полный либо сокращенный). Тестирование работы данного алгоритма на текстах НКРЯ показало, что при использовании сокращенного словаря точность определения позиции ударения уменьшается всего на 1,4 %.

Для хранения сокращенного словаря в памяти использовались хэш-коды слов и коды варианта позиции ударения, что позволило уменьшить физический объем словаря до 790 кБ.

Словарь такого объема может использоваться в системах синтеза речи для мобильных устройств, характеризующихся малым объемом памяти.

Авторы благодарят разработчиков Национального корпуса русского языка, позволившего провести данные исследования, и персонально Леонида Лейбовича Иомдина и Алексея Игоревича Зобнина за помощь в вычислении частоты встречаемости слов в НКРЯ.

Список литературы

1. SMS2Voice. Сервис голосовых сообщений [Электронный ресурс]. – 2002. – Режим доступа : http://voice.s-soft.org. – Дата доступа : 08.12.11.

2. What Does Garmin GPS System Text to Speech Mean? // eHOW [Electronic recourse]. – 2010. – Mode of access : http://www.ehow.com/info_8550345_garmin-system-text-speech-mean.html. – Date of access : 08.12.11.

3. VoiceOver // Apple [Электронный ресурс]. – 2011. – Режим доступа: http://www.apple.com/ru/accessibility/voiceover/. – Дата доступа : 08.12.11.

4. Acapela TTS for Windows Mobile // Acapela [Electronic recourse]. – 2005. – Mode of access : http://www.acapela-group.com/acapela-tts-for-windows-mobile-2-2-speech-solutions-tts.html. – Date of access : 08.12.11.

5. Nuance TALKS // Nuance [Electronic recourse]. – 2004. – Mode of access : http://www.nuance.com/talks/. – Date of access : 08.12.11.

6. Mobile Speak [Electronic recourse] // Code Factory. – 2004. – Mode of access : http://www.codefactory.es/en/products.asp?id=316. – Date of access : 08.12.11.

7. Gartner Says Worldwide Mobile Phone Sales Grew 17 Per Cent in First Quarter 2010. Press Release // Gartner [Electronic recourse]. – 2010. – Mode of access : http://www.gartner.com/it/page.jsp?id=1372013. – Date of access : 08.12.11.

8. Gartner Says Sales of Mobile Devices Grew 5.6 Percent in Third Quarter of 2011; Smartphone Sales Increased 42 Percent. Press Release // Gartner [Electronic recourse]. – 2011. – Mode of access : http://www.gartner.com/it/page.jsp?id=1848514. – Date of access : 08.12.11.

9. Цирульник, Л.И. Алгоритмы создания и пополнения грамматического словаря русского языка для синтеза речи по тексту / Л.И. Цирульник, В.В. Веремей // Информатика. – 2012. – № 1 (31). – С. 27–38.

10. Алгоритм поиска ударений [Электронный ресурс]. – 2007. – Режим доступа : http://proteus2001.narod.ru/802/data/00.htm#accent search. – Дата доступа : 08.12.11.

11. Автоматическое определение места ударения в незнакомых словах в системе синтеза речи / О.Г. Хомицевич [и др.] // Материалы XXXVI Междунар. филологической конф. – СПб., 2008. – С. 175–183.

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

13. Национальный корпус русского языка [Электронный ресурс]. – 2000. – Режим доступа : http://www.ruscorpora.ru/. – Дата доступа : 08.12.11.

14. Кнут, Д.Э. Искусство программирования / Д.Э. Кнут. – 2-е изд. – М. : Вильямс, 2011. – Т. 3 : Сортировка и поиск. – 824 с.

Поступила 08.12.2011

Объединенный институт проблем
информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail: liliya.tsirulnik@gmail.com,
dima.pokladok@gmail.com

L.I. Tsirulnik, D.A. Pokladok

GRAMMATICAL DICTIONARY AND STRESS PLACEMENT RULES FOR MOBILE DEVICE TTS-SYNTHESIS

Studies of stress placement regularities for Russian language are described. The research shows that the size of the dictionary can be significantly reduced. An algorithm of word stress placement is suggested. Testing of the results shows that the size of the dictionary can be reduced eight times, while the word stress placement accuracy decreases just 1.4 %. Reducing dictionary size allows using it in mobile devices, which are characterized by a limited memory.