Программа анализа фонетических статистик в текстах на русском языке и ее использование для решения прикладных задач в области речевых технологий

Software for automated statistical analysis of phonetic units frequency in Russian texts and its application for speech technology tasks

Смирнова Н. С. (nsmirnova@speechpro.com)
Чистиков П. Г. (chistikov@speechpro.com)

Центр речевых технологий (http://speechpro.ru)

В статье представлена реализация статистического анализатора текста TextAnalyser, описаны основные возможности его использования в сфере речевых технологий и приведены некоторые результаты статистической обработки в анализаторе большого текстового корпуса, в частности, ранжированные по частотности списки аллофонов русских фонем и наиболее частотных бифонемных сочетаний. Программа TextAnalyser и получаемые с ее помощью статистические данные могут быть полезны при разработке систем автоматического синтеза и распознавания речи.

Введение

Многие современные направления разработки речевых технологий связаны с использованием речевых баз данных, полученных на основе текстов в произнесении одного или нескольких дикторов. Критерием пригодности речевой же базы служит, прежде всего, полнота представления в ней речевых элементов, избранных в качестве основных или релевантных для решения тех или иных прикладных задач. Так, например, для разработки систем синтеза речи состав таких элементов может быть различным в зависимости от типа избранной базовой единицы — чаще всего это дифоны или аллофоны (трифоны). Для дифонного синтеза требуется база данных, содержащая все возможные для заданного языка двучленные комбинации фонем (аллофонов), тогда как при аллофонном синтезе необходим учет всех возможных сочетаний левого и правого контекста (как правило, различные типы контекстов при этом объединяются в классы по степени акустической близости, чтобы сократить общий инвентарь единиц). Аналогичный подход применяется и при разработке систем распознавания

речи, построенных на контекстно зависимых монофонах (трифонах). В случае разработки мультимедийных справочных экспертных систем принципиальное значение будет иметь наличие в речевом материале элементов, способствующих проявлению акцентной/диалектной вариативности речи на заданном языке, а также междикторской регионально не обусловленной вариативности.

В связи с изложенным выше, особое значение приобретает подготовка текстового материала в задачах, где набор необходимых элементов речевой базы заранее определен, особенно в тех случаях, когда ресурс для обработки и структурирования речевого материала ограничен. Помимо фонетической представительности, использование специального текстового материала фиксированного объема обеспечивает компактность, что в дальнейшем позволяет существенно сократить время на его обработку. Считается, что при использовании больших объемов текстового материала полнота покрытия единиц достигается и без специального подбора, однако, во-первых, при таком подходе неизбежно возникает избыточность базы (которая может в дальнейшем потребовать пост-обработки материала для исключения повторяющихся элементов) и, во-вторых, отдельные редкие фонетические сочетания, возникающие на стыках слов, вполне могут так ни разу и не встретиться. Именно поэтому даже в сегодняшних синтезаторах, основанных на технологии Unit Selection и имеющих базы данных объемом более 10 часов речи, изредка встречаются «пропуски» или не вполне адекватные замены — по причине отсутствия необходимых элементов в исходной акустической базе.

С целью упрощения и автоматизации процесса формирования текстов с заданной фонетической представительностью, оценки степени полноты покрытия элементов в заданном тексте, а также сокращения избыточности текстового корпуса за счет удаления из него повторяющихся элементов, был разработан фонетический анализатор текстового материала, описание которого приводится ниже.

Описание анализатора частотности фонетических единиц в текстовом материале TextAnalyser

TextAnalyser позволяет решать следующие задачи:

- анализ статистики встречаемости в тексте речевых единиц различных уровней: фонем, аллофонов, слогов, звуковых последовательностей, слов;
- генерация слов (последовательностей слов), содержащих фонетические единицы с заданными параметрами;
- оценка степени фонетической информативности слов, входящих в состав текстового материала.

В состав программной системы входят следующие основные части:

- словарь словоформ русского языка, объёмом 180 000 словоформ [1]

- автоматический фонетический транскриптор русской речи по тексту, разработанный для синтезатора русской речи [2].
- модуль статистической обработки транскрипционного материала, осуществляющий сегментацию затранскрибированного материала на фонетические единицы различных типов — аллофоны, двучленные и трехчленные звуковые последовательности, слоги (выделенные по заданным правилам), — и сравнение полученной статистики с опорной статистикой или «нормой» (см. описание ниже), а также получение статистики встречаемости в тексте произвольно заданных типов звуковых последовательностей; поиск в словаре и вывод слов или словосочетаний, содержащих заданные аллофоны, звуковые последовательности или слоги.

«Норма» встречаемости в тексте единиц различных типов определяется на основе статистического анализа фонетически представительного текстового корпуса. В текущей версии анализатора для расчета опорной статистики используется словарь объемом в 460 965 слов (2 500 288 фонемоупотреблений).

Поскольку основной произносительной единицей принято считать слог, для оценки слоговой представительности текстов в программе предусмотрено пять альтернативных вариантов слогоделения с последующим вычислением статистики встречаемости слогов:

1) Деление на открытые слоги (слоги всегда оканчиваются на гласный).
2) Деление на закрытые слоги (слоги оканчиваются на согласный, если за гласным следуют два или более согласных).
3) Выделение в тексте последовательностей типа «согласный + гласный» (при этом остальные элементы не учитываются).
4) Деление на слоги по правилу Р. И. Аванесова [3].
5) Деление на слоги по правилу Л. В. Щербы [4].

При этом слоговая статистика может сниматься как в «фонемном» (т. е. без учета ударности/безударности и редукции гласных), так и в «аллофонном» представлении, с полноценным учетом всех позиционных и комбинаторных аллофонов.

Кроме слоговой статистики, программа позволяет вычислять статистику встречаемости звуков и звуковых последовательностей в следующих вариантах:

- монофоны (статистика для каждого аллофона без учета контекста);
- дифоны (статистика для последовательностей из двух аллофонов, т. е. для каждого аллофона с раздельным учетом правого и левого контекстов);
- трифоны (статистика для последовательностей из трёх аллофонов, т. е. для каждого аллофона с учетом левого и правого контекстов).

Все перечисленные виды статистической оценки могут использоваться для произвольно выбранного текста.

Кроме того, может быть получена сравнительная оценка встречаемости в тексте фонетических единиц относительно аналогичной статистики

в опорном тексте, принятом за «норму» (который также может быть произвольно задан пользователем). Это позволяет определить, каких элементов не хватает в тестовом материале, и при необходимости дополнить текст, используя приведенные примеры слов и словосочетаний с требуемыми фонетическими единицами. Слова, в которых встречаются искомые фонетические единицы, выводятся в модифицированной орфографии (промежуточный уровень между нормативным орфографическим написанием и фонетической транскрипцией). Ударные гласные передаются прописными символами. Пример использования данной функции приведен на Рис. 1:

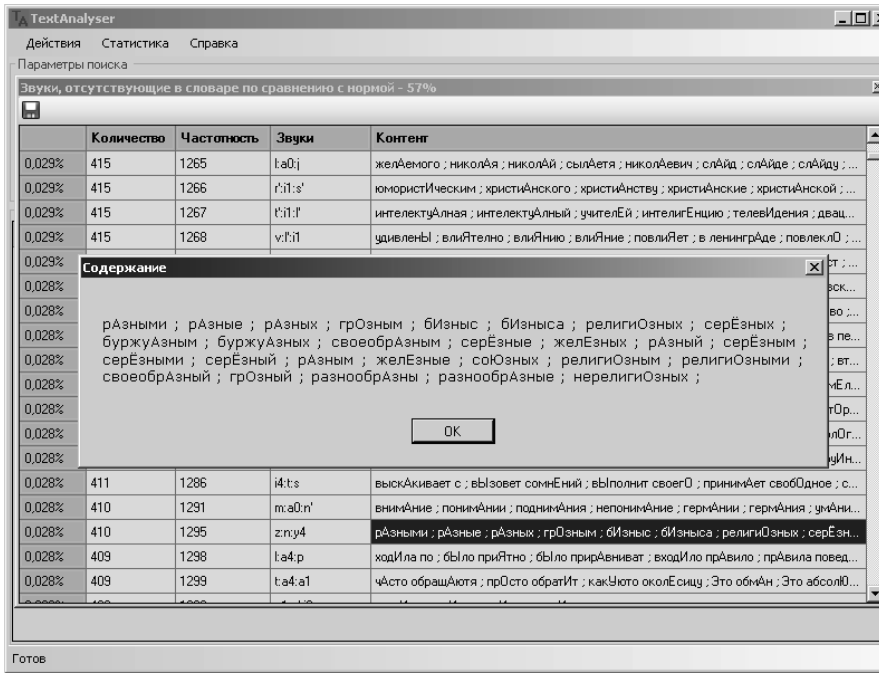

Рис. 1. Отображение типов слога (выделенных по правилам Р. И. Аванесова), отсутствующих в тексте, с примерами слов из опорного текстового корпуса, которые содержат заданный тип слога

Опция вывода статистики информативности слов позволяет оценить, насколько часто повторяются в словах текстового материала заданные фонетические единицы (аллофоны, дифоны, трифоны, слоги). Если какой-либо элемент заданного типа встречается в тексте дважды или чаще, то каждому слову, его включающему, сопоставляются те слова, в которых данный компонент дублируется.

Кроме того, для входящих в слово компонентов указывается их ранг по частотности в опорном словаре-норме (соответственно, на Рис. 2 это «39», «69» и «133» для слогов слова «доро́ги» при слогоделении «по Аванесову»):

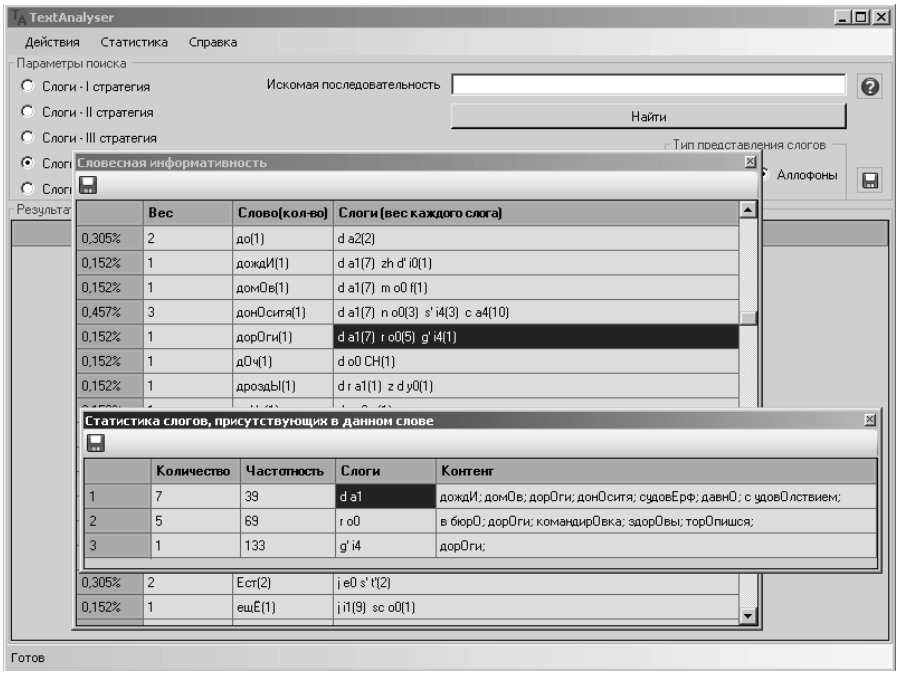

Рис. 2. Оценка информативности слов по выбранной единице анализа (слоги)

Имеется возможность сортировать наборы элементов по минимальному порогу встречаемости в тексте (например, более двух раз, более трех раз и т. п.).

Кроме расчета статистики фонетических единиц данное приложение предусматривает также следующие возможности:

- сбор статистики встречаемости слов в тексте (частота встречаемости каждой словоформы);
- учет (исключение) повторяющихся словоформ при выведении текстовых последовательностей (слов) с заданным сочетанием фонем (аллофонов);

Некоторые статистические данные, полученные с помощью TextAnalyser

Сегодня исследование статистических характеристик речевых единиц является крайне востребованным и актуальным уже не столько в теоретических целях, сколько для эффективного решения конкретных прикладных задач, касающихся обработки и моделирования речевого сигнала (в частности, в вероятностных n-gram алгоритмах автоматического распознавания речи и др.). Поэтому в данном разделе мы приводим некоторые статистические данные, полученные с помощью программы анализатора статистик TextAnalyser, и частично сопоставляем их с ранее публиковавшимися сведениями.

Для вычисления статистических характеристик русского языка был сформирован текстовый корпус, включающий в себя примерно в равных пропорциях тексты из классической русской литературы (фрагменты произведений Ф. М. Достоевского, А. П. Чехова, Н. В. Гоголя, М. Ю. Лермонтова), современную русскую прозу (фрагменты произведений В. Г. Распутина, О. Михайлова, И. С. Шмелева, В. А. Солоухина, В. Н. Крупина, Ч. Айтматова и др.) и публицистику (опубликованные новостные репортажи, интервью, текстовые расшифровки дискуссий и публичных лекций). Общий объем текстового материала составил более 460 тысяч (460 965) словоформ, более 1 млн. слогов, более 2,5 млн. (2 500 288) фонемоупотреблений. Учитывались все аллофоны фонем русского языка, включая предударные и заударные гласные, а также позиционные и комбинаторные аллофоны согласных не только внутри, но и на стыках слов.

На вход программы был подан текстовый материал в орфографической форме, после чего он был подвергнут автоматическому членению на синтагмы и транскрибированию в модуле транскриптора системы синтеза речи. На базе данного текстового корпуса была получена статистика реализации в русских текстах всех типов фонетических единиц, предусмотренных для выделения в TextAnalyser.

В таблице 1 приведены данные относительно встречаемости в тексте аллофонов русских фонем:

Таблица 1. Статистика аллофонов русских фонем в большом текстовом корпусе

Аллофон	Количество	Ранг	Аллофон	Количество	Ранг
a1	136 247	1	**a2**	50 346	20
a4	134 417	2	**d**	46 516	21
i4	129 413	3	**i0**	45 618	22
i1	123 175	4	**t’**	44 941	23
a0	107 527	5	**y4**	42 162	24
t	107 481	6	**s’**	39 080	25
j	104 064	7	**r’**	35 575	26
n	93 916	8	**z**	33 011	27
o0	90 216	9	**ch**	29 413	28
s	85 979	10	**b**	28 764	29
v	77 550	11	**sh**	28 388	30
r	75 458	12	**u1**	28 340	31
e0	73 723	13	**g**	27 903	32
k	72 485	14	**u4**	27 670	33
n’	59 279	15	**u0**	27 620	34
p	57 808	16	**d’**	27 504	35
m	57 748	17	**f**	25 831	36
l’	53 374	18	**v’**	24 539	37
l	52 090	19	**zh**	23 804	38

Аллофон	Количество	Ранг	Аллофон	Количество	Ранг
m’	22 893	39	**g’**	4 385	50
y0	22 154	40	**f’**	2 368	51
h	21 715	41	**H**	1 748	52
c	20 400	42	**h’**	1 076	53
y1	17 383	43	**o4**	222	54
k’	11 502	44	**CH**	133	55
p’	11 267	45	**C**	86	56
sc	9 819	46	**e1**	12	57
b’	8 809	47	**SC**	11	58
o1	8 734	48	**e4**	9	59
z’	6 587	49			

В данной системе транскрипции для аллофонов русских гласных используются символы <a> (а), <i> (и), <e> (е,э), <u> (у), <o> (о), <y> (ы); индекс «0» обозначает ударный гласный, «1» — предударный (или первый предударный аллофон фонемы / а /), «2» — второй предударный аллофон фонемы / а /, «4» — гласный в заударном слоге. Значок «‘» обозначает мягкость согласного. Приведем несколько примеров слов, записанных в данной системе транскрипции:

- город: g o0 r a4 t
- карандаш: k a2 r a1 n d a0 sh
- зонтик: z o0 n’ t’ i4 k

Как видно из таблицы, наиболее частотными являются безударные аллофоны фонем /а/ и /i/. У фонем /u/ и /y/, не подвергающихся качественной редукции в безударном положении, частотность безударных аллофонов также выше, чем ударных.

Таблица содержит также аллофоны, возникающие исключительно на стыках слов в результате процессов ассимиляции. В частности, это озвонченные аллофоны непарных глухих фонем / h /, / ch /, / c / и / sc / — соответственно / H /, / CH /, / C / и / SC /. Они, как видно, встречаются довольно редко, однако их моделирование является необходимым в полноценных системах синтеза и распознавания речи.

В силу некоторых различий в правилах транскрибирования текстового материала полученные нами данные не могли быть полноценно сопоставлены с ранее опубликованными исследованиями [5, 6]. Для целей сопоставления информация о частотности аллофонов, полученная с помощью TextAnalyser, была сведена к фонемам и таким образом оказалось возможным провести сравнение наших данных со статистикой, представленной в [6]. Результаты приведены в таблице 2.

Данные ЦРТ, полученные на материале 2,5 млн. фонемоупотреблений, сопоставляются с данными Института математики (ИМ) Сибирского отделения АН СССР (1964 г., на материале 1,5 млн. фонемоупотреблений в текстах технической и математической тематики) и статистикой, полученной в Лаборатории

экспериментальной фонетики ЛГУ (ЛЭФ) на материале 100 000 фонемоупотреблений. В сравнительной таблице приводятся только ранги фонем, поскольку анализировались текстовые выборки разного объема.

Принятая для представления «ленинградская» трактовка фонемного статуса аллофонов «ы» и «а» в безударной позиции на месте орфографического «о» объясняется исключительно практическим удобством и не связана с определенной теоретической позицией авторов статьи.

Таблица 2. Сопоставление статистики русских фонем в различных исследованиях

Фонема	Ранг			Фонема	Ранг		
	ЦРТ	ЛЭФ	ИМ		ЦРТ	ЛЭФ	ИМ
a	1	1	1	**r’**	22	21	22
i	2	2	2	**z**	23	22	23
t	3	3	4	**ch**	24	24	25
j	4	6	5	**b**	25	25	24
o	5	4	6	**sh**	26	26	32
n	6	5	3	**g**	27	27	28
s	7	7	7	**d’**	28	30	27
u	8	8	9	**f**	29	28	29
y	9	15	20	**v’**	30	31	30
v	10	11	10	**zh**	31	29	33
r	11	9	8	**h**	32	32	31
e	12	13	12	**m’**	33	33	26
k	13	10	11	**c**	34	34	34
n’	14	14	16	**k’**	35	36	36
p	15	16	14	**p’**	36	37	35
m	16	17	15	**sc**	37	35	37
l’	17	18	13	**b’**	38	38	39
l	18	12	18	**z’**	39	39	38
d	19	19	21	**g’**	40	40	40
t’	20	20	17	**f’**	41	41	41
s’	21	23	19	**h’**	42	42	42

Как видно из таблицы, несмотря на различия в объемах и стилистической принадлежности текстовых выборок, в целом данные всех трех источников очень близки. Совпадает состав первых наиболее частотных семи фонем — с разницей в рангах не более двух (a, i, t, j, o, n, s) и наиболее редких девяти (c, k’, p’, sc, b’, z’, g’, f’, h’). Некоторые расхождения с одним из источников присутствуют в рангах фонем / j /, / l / и / m’ /. Примечательно, что в данных ЦРТ ранг / j / занимает как бы промежуточное положение по сравнению с рангом этой фонемы по ЛЭФ и ИМ, ранг / l / совпадает с данными ИМ при существенном

различии в ранге этой фонемы у ЛЭФ, тогда как ранг / m' / совпадает с данными ЛЭФ при существенных различиях с данными ИМ.

Наиболее существенные различия наблюдаются в ранге фонемы / y /(«ы»): 9 ранг у ЦРТ по сравнению с 15 у ЛГУ и с 20 по данным ИМ. Т.е. разница между ЦРТ и ИМ составляет 11 рангов (!). Не имея достаточно подробных сведений о специфике обработки текста исследователями ЛЭФ и ИМ, мы можем лишь предположить, что причина данного различия может скрываться в характере транскрибирования начальнословных «и» после слов, оканчивающихся на твердый согласный («как **И** ты», «в **И**нтересах», «всех **И**х» и т. п.). В нашем случае они всегда транскрибировались как аллофоны «ы» (если модуль автоматического разбиения на синтагмы не ставил между ними синтагматической границы). Если же причислять их к аллофонам фонемы «и», то ранг фонемы «ы» понизился бы до 14, что близко к данным ЛЭФ и ИМ.

Приведем еще одну таблицу статистических данных — а именно, бифонемных сочетаний. Информация о типах и частотности таких сочетаний необходима для обеспечения полноты покрытия единиц в системах дифонного синтеза речи.

Статистика построена на анализе уже упомянутого выше текстового корпуса. Всего в тексте было обнаружено более 2 тыс. типов таких последовательностей (2176). Из них 2012 последовательностей встретились более 3 раз.

Наиболее частотные 164 последовательности, составляющие 50 % всех бифонемных сочетаний, приведены в таблице 3.

Таблица 3. Статистика дифонов — последовательностей из двух аллофонов

Дифон	Ранг	Количество	Дифон	Ранг	Количество
j:i4	1	27 609	**l:a4**	18	13 081
n:a4	2	22 845	**t':i4**	19	13 058
a4:_	3	22 647	**p:a1**	20	11 680
s:t	4	22 511	**y4:j**	21	11 451
i4:j	5	19 963	**a0:j**	22	11 321
i4:_	6	18 551	**n:a1**	23	11 086
n':i1	7	18 527	**j:u4**	24	11 066
n':i4	8	18 382	**a4:v**	25	11 044
v:a4	9	17 839	**a1:t**	26	10 820
r':i1	10	17 040	**j:a4**	27	10 701
a4:j	11	16 383	**n:y4**	28	10 567
l':i4	12	15 512	**p:a2**	29	10 404
t:a4	13	15 484	**v:a1**	30	10 314
r:a1	14	15 072	**n:a0**	31	10 270
k:a1	15	14 522	**r:a0**	32	10 252
k:a4	16	14 223	**t:a1**	33	10 232
t:o0	17	13 433	**p:r**	34	9 954

Дифон	Ранг	Количество	Дифон	Ранг	Количество
v:o0	35	9954	**_:s**	77	6517
s:k	36	9854	**r:a2**	78	6516
j:i1	37	9821	**a1:r**	79	6507
e0:t	38	9677	**u4:_**	80	6465
_:a1	39	9203	**zh:y4**	81	6454
a1:s	40	9142	**m':i1**	82	6436
p:r'	41	9130	**l':i1**	83	6381
s':t'	42	9051	**a0:t'**	84	6375
_:p	43	9032	**o0:r**	85	6363
t:a0	44	8919	**t':i1**	86	6299
i1:s	45	8896	**i1:z**	87	6266
a4:m	46	8832	**m:_**	88	6216
i1:v	47	8690	**sh:y4**	89	6168
j:a0	48	8655	**o0:t**	90	6143
e0:n'	49	8645	**m:a1**	91	6115
k':i4	50	8572	**t:r**	92	6065
_:i1	51	8552	**a4:s'**	93	6030
a0:l	52	8426	**i1:r**	94	6023
v:a0	53	8363	**d:a1**	95	5891
d:a0	54	8183	**t:_**	96	5890
_:k	55	8141	**k:o0**	97	5863
r:o0	56	8085	**n:o0**	98	5863
n:a2	57	8076	**f:s'**	99	5858
d':e0	58	7947	**i4:m**	100	5858
sh:t	59	7508	**v':i4**	101	5819
a1:v	60	7428	**a1:j**	102	5787
m':i4	61	7385	**i4:s**	103	5785
c:a4	62	7359	**j:_**	104	5782
i4:n	63	7329	**o0:m**	105	5782
a1:k	64	7311	**l':e0**	106	5781
ch:i4	65	7141	**e0:j**	107	5765
o0:n	66	7128	**a1:b**	108	5749
n':e0	67	7029	**ch:i1**	109	5677
s':i1	68	6979	**r':e0**	110	5675
_:n	69	6908	_:v	111	5605
v':i1	70	6899	a1:d	112	5593
i4:t	71	6898	_:sh	113	5576
v':e0	72	6830	a0:l'	114	5549
k:a0	73	6782	a0:t	115	5546
r:a4	74	6748	i1:n	116	5529
l:a1	75	6611	l:a0	117	5501
o0:j	76	6589	a1:g	118	5459

Дифон	Ранг	Количество	Дифон	Ранг	Количество
a0:s	119	5 411	a1:z	142	4 884
p':i1	120	5 384	d:a4	143	4 883
d':i4	121	5 308	i1:k	144	4 871
n':i0	122	5 264	a4:p	145	4 866
m':e0	123	5 262	a1:n	146	4 811
t:o1	124	5 219	o0:v	147	4 787
t:v	125	5 196	i1:m	148	4 686
i1:n'	126	5 119	a0:_	149	4 668
i1:t	127	5 106	a2:s	150	4 626
a1:n'	128	5 092	i1:r'	151	4 582
a4:f	129	5 077	a4:n	152	4 532
z:a1	130	5 034	a1:d'	153	4 522
i4:v	131	5 013	s:p	154	4 479
t':i0	132	4 995	e0:s	155	4 471
l':i0	133	4 964	a1:l'	156	4 434
a4:s	134	4 961	a0:m	157	4 432
d':i1	135	4 955	a0:k	158	4 419
a0:n'	136	4 953	r':i4	159	4 412
o0:l'	137	4 945	i1:p	160	4 390
a0:n	138	4 932	l':n	161	4 384
y4:_	139	4 921	e0:r	162	4 380
_:j	140	4 915	j:e0	163	4 379
i4:l'	141	4 903	b:y0	164	4 376

Аналогичным образом была получена статистика трифонов (всего 35 073 типа, из них 24 342 — трифоны с частотой встречаемости в опорном тексте более 3 раз), последовательностей «согласный плюс гласный» (всего 410 типов) и слогов, выделенных по различным правилам. Любопытно, что число типов слога при слогоделении по Л. В. Щербе на несколько сот превышает число слогов, выделенных по правилу Р. И. Аванесова: соответственно 11 354 и 10 801, при этом в обоих случаях лишь чуть больше половины слогов (6164 и 5841) имеют частоту встречаемости в текстовом корпусе более трёх раз. В силу ограниченного объема данной публикации мы не имеем возможности приводить полученные статистические данные в полном объеме.

Перспективы развития TextAnalyser

Дальнейшая разработка программы предполагает автоматизацию процесса формирования текстового материала с заданной представительностью фонетических единиц (фонем, дифонов, трифонов, слогов). Это необходимо, в частности, при разработке систем автоматического синтеза речи. Так,

используя данную программу, появится возможность получить текстовый материал с полным покрытием всех возможных в русском языке дифонов, или, например, обеспечить покрытие нескольких сот наиболее частотных типов слога. Разумеется, речь не идет о генерации оригинального связного текста — такое машине пока не под силу. Имеется в виду компоновка текста из подходящих слов, словосочетаний и предложений, содержащихся в опорном текстовом корпусе. Подбор их будет осуществляться программой таким образом, чтобы минимизировать избыточность (повторяемость) единиц в тексте. Критерием могут служить не только фонетические (сегментные) свойства текста, но и коммуникативно-синтаксическая составляющая (по знакам препинания): например, можно будет задать долю вопросительных предложений, или предложений, содержащих неконечные синтагмы (по запятым). Очевидно, что эффективность подбора зависит от параметров опорного текстового корпуса: чем более полным и разнообразным он будет, тем более информативным и компактным будет получаемый на его основе текстовый материал.

Литература

1. *Хомицевич О. Г., Рыбин С. В., Таланов А. О., Опарин И. В.* Автоматическое определение места ударения в незнакомых словах в системе синтеза речи // Материалы XXXVI Международной филологической конференции. Санкт-Петербург, 2008.
2. *Вольская Н., Коваль А., Коваль С., Опарин И., Погарева Е., Скрелин П., Смирнова Н., Таланов А.* Синтезатор русской речи по тексту нового поколения // Материалы XXI международной конференции Диалог 2005. Москва, 2005.
3. *Аванесов Р. И.* О слогоразделе и строении слога в русском языке // Вопросы языкознания, № 6, 1954.
4. *Бондарко Л. В.* Фонетика современного русского языка. Санкт-Петербург, 1998. С. 199–200.
5. *Елкина В. Н., Юдина Л. С.* Статистика слогов русской речи // Вычислительные системы. Новосибирск, 1964. № 10, с. 58–62.
6. *Бондарко Л. В., Зиндер Л. Р., Штерн А. С.* Некоторые статистические характеристики русской речи // Слух и речь в норме и патологии. Ленинград, 1977. Вып. 2, с. 3–16.