УДК 003.23: 004.522

ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ КАЗАХСКОГО ЯЗЫКА ДЛЯ СИНТЕЗА РЕЧЕПОДОБНЫХ ПОМЕХ

О.Б. ЗЕЛЬМАНСКИЙ, Х.С. АБИШЕВ

Белорусский государственный университет информатики и радиоэлектроники
ул. П. Бровки, 6, г. Минск, 220013, Республика Беларусь
7650772@rambler.ru

На сегодняшний день статистические закономерности казахского языка не достаточно изучены. В настоящей работе представлены результаты исследования частоты употребления букв казахского алфавита и слов казахского языка на примере анализа казахско-русского словаря, технических и газетных текстов и их сравнительный анализ.

Ключевые слова: статистика казахского языка, частота употребления букв казахского алфавита, встречаемость слов казахского языка, речеподобные сигналы.

Компиляция участков естественной речи широко используется в системах синтеза речи, а также речеподобных сигналов, формируемых по случайному закону и по своим основным временным, спектральным характеристикам и восприятию на слух, максимально подобных речевым сигналам, но не содержащих смысловой информации. Один из предлагаемых способов синтеза речеподобных сигналов заключается в преобразовании формируемого с учетом статистических закономерностей выбранного языка орфографического псевдотекста в акустические колебания звукового диапазона частот. Таким образом, для реализации подобных систем необходимо формирование речевых баз дикторов, представляющих собой набор отрезков естественной речевой волны, соответствующих фонетическим единицам речи диктора. При этом необходим учет статистических закономерностей языка. Для русского, английского и арабского языков такие системы достаточно подробно изучены. Однако для казахского языка сведения о подобных системах не опубликованы.

Таким образом, с целью создания речевых баз, учитывающих основные статистические, фонетические, лексические и грамматические особенности современного казахского языка, был проведен статистический анализ употребления букв и слов казахского языка.

Казахский кириллический алфавит, разработанный С.А. Аманжоловым и принятый в 1940 году, содержит 42 буквы: 33 буквы русского алфавита и 9 специфических букв казахского языка Ә, Ғ, Қ, Ң, Ө, Ұ, Ү, Һ, І.

Для создания баз данных и возможной сегментации речевых сигналов на казахском языке предложена методика статистического учета употребления букв казахского алфавита. За основу методики оценки употребления той или иной буквы принимали их количество, использованное в казахско-русском словаре. При этом подсчитывалось количество слов, начинающихся с определенной буквы. В табл. 1 представлены результаты статистической обработки казахско-русского словаря на 50000 казахских слов, выборки из технических текстов, а так же статей средств массовой информации (СМИ) различной тематики.

Анализ табл. 1 показывает, что полученные статистические закономерности очень схожи между собой. Так, наиболее используемой буквой казахского алфавита является буква «Қ», за которой следует буква «Т», а наименее употребляемыми

являются буквы В, Ғ, Ц, Ч, Щ, Ю, Я. В казахском языке нет слов начинающихся с букв Ё, Ң, Й, Һ, Ъ, Ь.

Табл. 1. Результаты статистического анализа употребления букв казахского алфавита

№	Буква	Частота употребления (%)			№	Буква	Частота употребления (%)		
		Словарь	Технический текст	СМИ			Словарь	Технический текст	СМИ
1	Қ	14,5	12	11	22	Ф	1	0,01	0,1
2	Т	10,6	10	9,4	23	І	0,8	0,8	0,8
3	Ж	9,6	10	9,3	24	Г	0,6	0,1	0,1
4	Б	8,7	10,6	12,3	25	Л	0,6	0,3	0,3
5	А	8,5	8,5	8,2	26	Р	0,6	0,9	0,9
6	К	7,4	6,3	7,3	27	Х	0,6	1,3	1,1
7	С	6,4	8	7,5	28	Э	0,6	0,6	0,6
8	Ш	4,8	1,5	2,3	29	У	0,6	0,6	0,5
9	М	3,7	5	4,4	30	В	0,3	0,1	0,1
10	Е	3,5	5	4,8	31	Ғ	0,3	0,5	0,3
11	Д	3,2	3,7	4,3	32	Ц	0,1	0,03	0,04
12	О	1,9	3,5	3,8	33	Ч	0,04	0,01	0,04
13	Ұ	1,6	1,7	1,6	34	Щ	0,02	0,01	0,04
14	Ө	1,3	2,6	2,5	35	Ю	0,02	0,01	0,02
15	Ү	1,2	1,6	1,4	36	Я	0,02	0,03	0,06
16	Ә	1,2	1,3	1	37	Ё	0	0	0
17	П	1,2	0,8	1	38	Ң	0	0	0
18	И	1,2	0,7	0,8	39	Й	0	0	0
19	Н	1,2	1	1	40	Һ	0	0	0
20	Ы	1,1	0,2	0,3	41	Ъ	0	0	0
21	З	1	0,7	0,8	42	Ь	0	0	0

Поскольку системы синтеза речеподобных сигналов ориентированы на повседневный светский казахский язык, то для них больше подходит распределение вероятности появления букв алфавита, полученное в ходе анализа статей СМИ. Данное распределение вероятности является основанием для выбора набора слов и словаря для создания речевых баз казахского языка. Речевая база будет содержать минимальное количество слов, начинающихся с наименее употребимой буквы Я и максимальное количество слов, начинающихся с буквы Қ. Таким образом, текст, содержащий, например, 10 слов, начинающихся с буквы «Я», потребует около 1833 слов, начинающихся с буквы Қ.

На основании анализа современных текстов на казахском языке (статей СМИ объемом 20000 слов) была определена относительная встречаемость казахских слов различной длины (табл. 2).

Табл. 2. Частота употребления в казахской речи слов различной длины

Число слогов в слове	1	2	3	4	5	6
Вероятность появления слова, %	19	23,1	30,3	17,7	6,3	3,6

Анализ табл. 2 указывает на преимущественное использование в казахском языке трехсложных слов (до 30,3 %) и двухсложных слов (до 23,1 %). Ударение в казахском языке всегда падает на последний слог. Если слово склоняется и к нему прибавляется окончание, то ударным становится прибавленное окончание.