

Распределение фонем в татарском тексте: опыт квантитативного исследования

Галиева Альфия Макаримовна, ООО «НекстГИС» (Москва);
amgalieva@gmail.com

Наряду с качественным описанием фонем, представленным в татарских грамматиках, требуется их квантитативное исследование, выполненное на репрезентативных выборках текстовых данных. В данной работе анализируется распределение фонем и фонемных сочетаний в повести Амирхана Еники «Әйтелмәгән васыятъ» («Невысказанное завещание»). Приводятся абсолютные и нормализованные частоты (с использованием квантилей стандартного нормального распределения) для каждой из 34 фонем, выделенных в тексте. Использованный подход к нормализации данных (вычисление z -статистик) основан на учете сведений как об общем количестве употреблений каждой фонемы, так и о распределении фонем в зависимости от их позиции в словоформе (в начальной, срединной и финальной позиции), что позволяет не терять информации о позиционных особенностях низкочастотных фонем. Величина отклонения z -статистик от квантилей стандартного нормального распределения, находящихся в интервале $(-1,96, 1,96)$, свидетельствует о степени упорядоченности, характере системной организации фонем в языке.

Синтагматическая составляющая фонологической системы языка, так же, как и инвентарь фонем, является уникальной для каждого языка. Поэтому далее исследуется сочетаемость татарских фонем на примере двух- и трехфонемных сочетаний. Квантитативная фонологическая информация дополняется грамматической интерпретацией; в частности, показано, что особенности распределения частотных типов фонемных сочетаний связаны со структурой словоформы и определяются агглютинативным строем татарского языка.

Ключевые слова: фонема, квантитативная лингвистика, татарский язык, сочетаемость фонем, фонология и грамматика

PHONEME DISTRIBUTION IN A TATAR TEXT: A QUANTITATIVE APPROACH

Alfiya M. Galieva, NextGIS Company (Moscow);
amgalieva@gmail.com

Along with the qualitative description of phonemes which is presented in Tatar grammars, their quantitative description is often required. Such a research should be performed on a sufficient volume of text data which would contain lexical items in their natural surroundings and reflect basic regularities of the language system.

This paper analyzes the distribution of phonemes and phoneme combinations in the story by Amirkhan Yeniki “Äytmägän vasiyät” (“The unspoken testament”). The research material was the text converted to a phonologically relevant form. The absolute and relative frequencies (using the quantiles of the standard normal distribution) are provided for each of the 34 phonemes distinguished in the text. The approach to data normalization (z -statistics) is based on information about both the number of phonemes depending on their position (wordinitial, wordfinal or wordmedial), and the total number of occurrences of each phoneme in the text; this allows us to retain information about positional features of low frequency phonemes and to find individual tendencies for each cell in the contingency table. The deviation of the z -statistics from the quantiles of the standard normal distribution $(-1.96, 1.96)$ indicates the character of phoneme sequences in the language. It is shown that for the absolute majority of Tatar phonemes, the z -statistics are far beyond the interval $(-1.96, 1.96)$, with the exception of borrowed phonemes $[u]$ and $[u]$, as well as the low-frequency phoneme $[x]$, which indicates the random nature of the distribution of these phonemes in Tatar.

The syntagmatic component of the phonological system, just like the inventory of phonemes, is unique for each language, so without such study the phonological description would be incomplete. That is why a special section examines Tatar phoneme sequences using the example of two and three phoneme combinations. Since there is often a gap in grammar books between the descriptions of the phonological and morphological levels, quantitative information on phonemes is complemented with grammatical interpretation; in particular, the author shows that high frequency combinations of phonemes are associated with affixes and stems, which is determined by the agglutinative structure of the Tatar language.

Key words: phoneme, quantitative linguistics, the Tatar language, combinations of phonemes, phonology and grammar

Введение

В значительной части классических и современных исследований по количественной лингвистике анализируются распределения языковых единиц по длине, структуре слога, распределения единиц текста и корпуса по частям речи и т. п., обосновываются и оттачиваются статистические процедуры анализа языковых данных. При этом эмпирические данные по языкам разных типов часто дополняются теоретическими ожидаемыми значениями, полученными при помощи той или иной модели. Так, для моделирования частот по длине слова используются различные модификации распределения Пуассона [Altmann 2013; Kromer 2001; Best 2008; Best 2011 и др.], закона Ципфа [Altmann 2002; Kromer 2002], закон Мензартата-Альтмана [Altmann 1980]; при моделировании частотности типов слоговых структур хорошие результаты дает использование функции Ципфа-Алексеева [Zörnig et al. 2019], распределения Ципфа-Мандельброта [Mandelbrot 1965; Radojičić et al. 2019; Galieva, Vavilova 2021] и т. п. В частности, журнал «Glottometrics» во многом специализируется на подобном моделировании единиц языка и текста на материале языков разных типов. Количественная лингвистика вносит существенный вклад в развитие теории языка как за счет внедрения новых теоретических конструктов, так и обеспечивая эпистемологический каркас на первый взгляд случайными эмпирическими данными, обнаруживая в них тенденции и закономерности.

Очевидно, что фонологическая система языка, блестяще описанная структуралистами и их последователями, представляющими разные традиции [Jakobson, Halle 1956; Jakobson, Waugh 1979; Chomsky, Halle 1968; Трубецкой 2000 и др.], является хорошим объектом для количественных исследований: текстовый инвентарь фонем поддается различным классификациям и статистическим оценкам, и неслучайно именно исследователи фонологической системы языка высказали мнение о неравномерном распределении фонем в зависимости от их позиции в словоформе [Байчур 1953; Щербак 1970; Jakobson, Waugh 1979; Altmann 2005 и др.].

Тем не менее исследований по количественной фонологии относительно немного, что легко объясняется тем, что фонематическая транскрипция требует большого объема рутинной ручной работы. Обобщенные данные по инвентарю фонем для 2186 языков мира можно получить в базе данных хранилища кросс-лингвистических фонологических данных PHOIBLE, разработанном в Институте эволюционной антропологии Макса Планка [Moran, McCloy 2019]. Инвентарь фонем стандартного татарского языка, представленный в репозитории PHOIBLE, включает 43 единицы.

Данные из репозитория PHOIBLE в дальнейшем могут стать материалом для количественных исследований. В частности, работа С. Брауна показывает, что, согласно данным 451 языка, распределение фонем положительно смещено (*positively skewed*) и может быть аппроксимировано при помощи логнормального распределения [Brown 2017]. При моделировании фонологических данных используются также регрессионные методы; так, в статье Г. Коломы регрессия применена к материалу 100 языков с четырьмя фонологическими переменными (количество согласных, количество гласных, различимость ударения, различимость тона). Предложенный метод основан на принципах синергетической лингвистики: автор говорит о том, что уравнения регрессии могут быть выведены путем решения задачи оптимизации, предполагающей, что уровни фонологических переменных в языке должны быть выбраны таким образом, чтобы максимально упростить декодирование и минимизировать затраты при звукопроизводстве [Coloma 2014].

Б. Кесслер использует количественный подход к исторической фонологии. Исследование включает в себя методы обнаружения и проверки звуковых соответствий, моделирования хода изменения фонем; производится вычисление наиболее вероятного генетического дерева, совместимого с набором нововведений, проверка значимости фонетических свидетельств для определения генетических взаимосвязей между языками и диалектами [Kessler 2015].

Количественные оценки фонологических явлений исследуются также в прикладном аспекте изучения второго языка, результаты таких исследований находят применение в совершенствовании методики преподавания неродных языков [Thomson 2016].

В настоящее время эмпирический материал по фонологии могут обеспечить размеченные речевые корпуса, которые представляют собой специализированные базы данных, состоящие из аудиофайлов и транскрипций текстов [Гришина 2005; Гришина, Савчук 2009; Bogdanova-Beglarian et al. 2017; Bogdanova-Beglarian et al. 2018 и др.]; именно данные из речевых корпусов часто становятся объектом фонетических и фонологических исследований [Turunen, Lipping 2008 и др.]. В статье Н. С. Смирновой и П. Г. Чистикова описаны основные возможности использования статистического анализатора текста TextAnalyser в сфере речевых технологий и приведены некоторые результаты статистической обработки большого текстового корпуса, в частности, представлены ранжированные по частотности списки аллофонов русских фонем и двухфонемных сочетаний [Смирнова, Чистиков 2011].

В татарских грамматиках [Татарская грамматика 1993; Татар грамматикасы 2015] фонемный уровень языка представлен без количественных данных, что легко объяснить тем, что на момент разработки грамматик отсутствовали достаточные массивы подготовленных фонологических данных. Тем не менее для татарского языка достаточно качественный материал для квантитативных исследований могут дать предварительно обработанные письменные тексты.

В данной работе анализируется распределение фонем в татарском тексте на материале повести Амирхана Еники «Эйтелмәгән васыять» («Невысказанное завещание»). Поскольку часто в грамматике языков имеется зазор между описаниями фонологического и морфологического уровней, квантитативные данные дополняются комментариями из области грамматики.

Краткие сведения о татарской фонологии

В академических грамматиках татарского языка говорится о том, что в современном татарском языке имеется 12 гласных фонем, 9 из которых — исконные: [a], [ə], [y], [ɣ], [ɨ], [e], [o], [ø], [u]. При этом выделяются 3 дополнительные гласные фонемы, встречающиеся в русских и европейских заимствованиях: [ō], [ë], [ē] [Татарская грамматика 1993: 88, Татар грамматикасы 2015: 43]. 9 исконных фонем и 3 отдельные фонемы для заимствованных слов выделяются и в работах ряда западных исследователей (например, [Berta 2015]). Долгота гласных не является дифференциальным признаком татарских фонем. Следует отметить, что существуют и другие подходы к выделению фонем в татарском языке: например, в работах В. А. Богородицкого [Богородицкий 1953], Н. А. Баскакова [Баскаков 1988], Г. Бурбиела [Burbiel 2018] выделяются 10 фонем; в частности, В. А. Богородицкий (а за ним и другие исследователи) дополнительно выделяет дифтонгоидный [ɨy] в качестве отдельной фонемы [Богородицкий 1953: 28].

Систему консонантизма образуют 25 исконных согласных фонем и 3 добавочные фонемы, употребляющиеся в заимствованиях: [ɟ], [ɟɟ], [ʁ] [Татарская грамматика 1993: 91; Татар грамматикасы 2015: 46].

Относительно позиционных вариантов фонем нужно отметить следующее: [a] в инициальных слогах произносится как [a^o]; в грамматиках для фонем [y], [ɣ], [u] в абсолютном конце слова приводятся позиционные варианты [y^w], [ɣ^w], [u^w]. Составителями грамматик отмечается, что бифонемность [y], [ɣ] в финальной позиции «доказана фонологическими, фонетическими и особенно экспериментально-фонетическими исследованиями» [Татарская грамматика 1993: 91; Татар грамматикасы 2015: 45].

В трудах В. А. Богородицкого, основавшего в 1884 году при Казанском университете первую в мире лабораторию экспериментальной фонетики, большое внимание уделялось вопросам изменения вокализма в зависимости от позиции гласных в словоформе, исследователь вскрывал тесную связь фонетики и фонологии с морфологией; к сожалению, многие идеи Богородицкого не находят отражения в современных академических грамматиках татарского языка, ориентированных на русские грамматики, в частности на «Русскую грамматику» 1980 года [Татар грамматикасы 2015: 43].

Бифонемная природа [y] и [ɣ] в позиции непервых слогов находит отражение и в исследованиях, вышедших за рубежом. Так, Н. Поппе говорит о том, что [y], [ɣ], [u] в исконных словах не встречаются в финальных и неинициальных слогах, будучи заменены на соответствующие алломорфы бифонемного происхождения [Porre 1963: 9]. На различия [y], [ɣ], [u] в позиции начального и неначальных слогов указывают и другие исследователи, например Г. Бурбиел [Burbiel 2018: 3].

Тезис о неравномерном распределении фонем в зависимости от их позиции в словоформе, высказанный в научной литературе, не только требует эмпирического подтверждения на текстовых данных, но и оценки, каковы эти смещения в ту или иную сторону, с использованием строгих статистических методов, поэтому основная часть данного исследования посвящена анализу распределения фонем в татарском тексте, причем за основу будут браться нормализованные значения и квантили нормального распределения.

Отбор языкового материала и подготовка данных

Часто в качестве материала для исследования фонологической или слоговой структуры слов того или иного языка берутся данные из лексикографических источников ([Altmann 2005; Мороз 2019]; и др.), что объясняется в первую очередь доступностью таких данных.

В данном исследовании мы не используем материалы татарских словарей, так как специфика отбора и представления форм слов в словарях во многом не согласуется с их реальным использованием в речи, что для татарского языка с его агглютинативной морфологией совершенно критично. Например, в словарях фиксируется большое количество заимствованных слов с нетипичной для татарского языка фо-

немной организацией, среди которых может оказаться очень большой процент редко используемых в текстах слов, что существенно искажает данные о распределении фонем. Кроме того, в словарях лексемы даются в основной форме (например, глаголы фиксируются в форме инфинитива или имени действия с их стандартными аффиксами), поэтому данные из лексикографических источников не отражают аффиксальных цепочек, которые связывают слова в предложениях, при этом они могут содержать непропорционально большое количество аффиксов инфинитива или имени действия. Все это приводит к тому, что распределения фонем, выделенных из единиц словника словаря, оказываются существенно смещены в ту или иную сторону (например, могут быть переоценены данные по относительно редким для текстов фонемам и фонемным сочетаниям и недооценены фонемы, типичные для аффиксов и аффиксальных цепочек). В данном исследовании нас интересует распределение фонем в словоформах в реальном употреблении, поэтому текстовый материал, представляющий образец классической татарской прозы XX века, мы считаем оптимальным. Этим обусловлен выбор эмпирической базы исследования — текста повести А. Еники «Невысказанное завещание».

Важным этапом в подготовке данных стало приведение кириллического текста к форме, близкой к фонологической: 1 символ — 1 фонема. Считается, что татарское письмо главным образом основано на фонетическом принципе, согласно которому слова пишутся так, как произносятся [Татарская грамматика 1993: 108]. Тем не менее имеется достаточно много разных случаев нарушения принципа 1 буква — 1 звук. Здесь нужно отметить 3 основных случая:

- непроизносимые символы (1 буква — 0 звуков), влияющие на произношение соседних букв: *ь* и *ъ*;
- буквы *е*, *ю*, *я* в определенных позициях (1 буква — 2 звука, причем ряд гласного определяется сингармонистическим классом слова, что приводит к вариантам);
- буквы *у* и *ү* после некоторых гласных (когда они произносятся как сонорный согласный (неслоговой гласный) *w*).

Кроме того, буква *в* может читаться как шумный согласный *[v]* (в словах, заимствованных из русского или европейских языков) или сонант *[w]* (в словах тюркского происхождения или в восточных заимствованиях) в зависимости от происхождения слов. В таблице 1 представлены примеры преобразования татарских словоформ из кириллической в фонологически релевантную форму.

Таблица 1. Примеры приведения словоформ к фонологическому виду

Графический облик словоформы	Транскрипция
<i>урман</i> 'лес'	урман
<i>ямьле</i> 'красивый, приятный'	jəmle
<i>аулау</i> 'охотиться'	awlaw
<i>кияу</i> 'жених, зять'	kiɟəw
<i>егет</i> 'молодой человек, юноша'	jeget
<i>ел</i> 'год'	jyl

Важная оговорка: выделение фонем в данной работе основано на письменном тексте и выполнено машиной (без последующей ручной проверки). Как следствие, в классе гласных фонем отдельно не выделены 3 гласные фонемы, встречающиеся лишь в русских и европейских заимствованиях: *[ɔ̃]*, *[ɛ̃]*, *[ɛ̃]*. В составе согласных фонем не различались следующие классы фонем: увулярные и неувулярные варианты для *[ɟ]* и *[k]*. Впрочем, следует отметить, что количество фонологически неассимилированных заимствований в анализируемом тексте относительно небольшое, что не должно, на наш взгляд, существенно исказить данные. С учетом этого были выделены 34 фонемы (см. таблицу 2), позиционные варианты фонем не рассматривались. Таким образом, наш материал не претендует на исчерпывающее описание фонологической системы татарского языка, а лишь представляет ее приближенное подобие, условную модель. Тем не менее эта модель позволяет получить вполне адекватные сведения о распределении и комбинаторных свойствах татарских фонем.

Распределение фонем в зависимости от позиции в словоформе

Приведем предварительные сведения об анализируемом тексте. Изначально (после токенизации) в тексте повести А. Еники содержалось 12063 слова (4283 уникальных словоформы), которые были записаны в фонологической транскрипции. На рисунке 1 представлено количество словоформ в зависимости от длины. Самыми частотными (2450 единиц) являются слова, состоящие из 5 фонем.

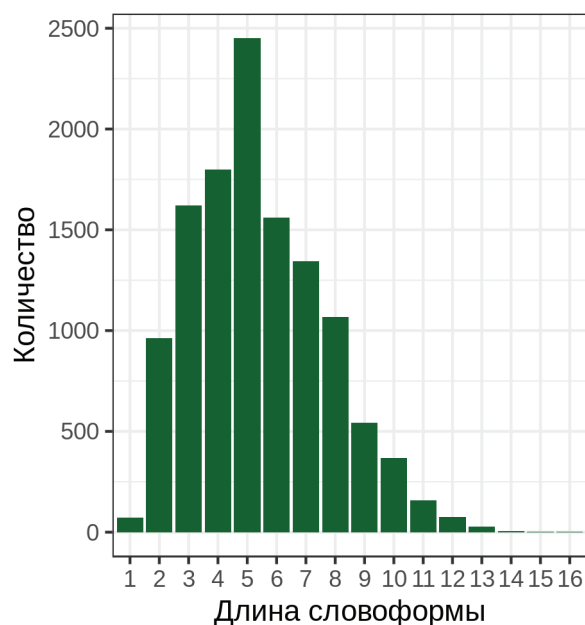


Рисунок 1. Количество словоформ в зависимости от их длины

Так как процедура анализа требовала подсчета числа фонем в начальной и финальной позиции и позиции в середине слова (сюда отнесены все фонемы за исключением крайних слева и справа), были удалены слова, состоящие из одной и двух фонем. В результате в анализируемом файле осталось 10944 словоформы (4202 уникальных словоформы), состоящие из 63102 фонем. Их распределение (абсолютные значения и z -статистики) в зависимости от позиции в словоформе представлены в таблице 2.

Таблица 2. Распределение фонем в зависимости от позиции в словоформе

Фонема	Значения в зависимости от позиции						Всего употреблений
	в начале слова		в конце слова		в середине слова		
	кол-во	Z	кол-во	Z	кол-во	Z	
а	881	-11,43	1396	5,76	4778	4,51	7055
э	374	-15,64	980	9,59	2969	4,81	4323
б	1510	69,89	0	-20,80	487	-39,05	1997
в	54	8,80	3	-4,05	53	-3,78	110
w	19	-5,33	36	-2,80	261	6,47	316
г	331	1,33	0	-19,65	1457	14,57	1788
д	265	-1,23	3	-18,57	1367	15,75	1635
е	89	-29,11	1152	13,42	3468	12,48	4709
ж	0	-1,02	0	-1,02	5	1,63	5
з	86	-7,89	255	6,01	708	1,49	1049
и	569	4,99	210	-13,58	1946	6,84	2725
j	601	14,22	356	-0,23	1118	-11,13	2075
к	2059	57,91	589	-5,01	1423	-42,08	4071
л	70	-25,63	183	-20,58	3457	36,75	3710
м	329	5,12	285	2,06	859	-5,71	1473
н	205	-24,47	1593	31,06	2911	-5,24	4709
ң	0	-12,66	470	32,79	285	-16,01	755
о	125	-2,13	10	-12,58	721	11,71	856
ө	146	7,11	0	-10,25	451	11,96	497
п	59	-11,63	1022	61,96	137	-40,03	1218
р	64	-26,83	990	13,49	2868	10,61	3922
с	626	23,59	41	-15,71	918	-6,27	1585
т	862	18,70	217	-13,99	1764	-3,75	2843
у	249	1,03	58	-12,83	1047	9,39	1354

Фонема	Значения в зависимости от позиции						Всего употреблений
	в начале слова		в конце слова		в середине слова		
	кол-во	z	кол-во	z	кол-во	z	
ү	210	8,50	20	-10,35	487	1,47	717
ф	18	-0,94	6	-3,76	103	3,74	127
х	188	20,13	0	-8,10	123	-9,57	311
ц	1	-1,32	0	-1,94	17	2,60	18
ч	275	8,13	99	-6,53	649	-1,27	1023
ш	321	9,52	72	-10,04	759	0,41	1152
щ	1	0,73	1	0,73	1	-1,16	3
ы	14	-31,15	894	4,96	3550	20,83	4458
ж	209	27,15	0	-7,36	48	-15,74	257
һ	133	22,43	2	-5,30	21	-13,62	156
Всего	10943	—	10943	—	41216	—	63102

Нормализованные значения z для фонем вычислены по методике, использованной в ряде исследований по квантитативной лингвистике [Altmann 2005: 193—194; Popescu et al. 2015: 32], когда значения для отдельной ячейки в таблице сопряженности вычисляются по формуле:

$$z = \frac{n_{ij} - E_{ij}}{\sqrt{\frac{n_{i\cdot} \cdot n_{\cdot j} (N - n_{i\cdot})(N - n_{\cdot j})}{N^2 (N - 1)}}},$$

где z — квантиль стандартного нормального распределения $N(0,1)$, n_{ij} — количество фонем в ячейке (i, j) ; $n_{i\cdot}$ — сумма в строке i ; $n_{\cdot j}$ — сумма в колонке j , а N — общее количество фонем в тексте. Выражение в знаменателе — стандартное отклонение для значения в ячейке (i, j) . Ожидаемое значение E для ячейки (i, j) вычисляется по формуле:

$$E_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{N}$$

Z-статистика, вычисленная по формуле, приведенной выше, удобна тем, что при оценке частотности фонемы в той или иной позиции в слове учитывается как суммарное количество употреблений данной фонемы в тексте, так и суммарное количество фонем в каждой позиции; таким образом, мы можем оценить индивидуальные тенденции для каждой ячейки в таблице сопряженности. Если бы фонемы в тексте были распределены случайным образом, z -статистики для них должны были бы располагаться в интервале квантилей стандартного нормального распределения $(-1,96, 1,96)$; следовательно, величина отклонения реальных z -статистик от данного интервала свидетельствует о степени упорядоченности, о характере системной организации фонем в языке.

В частности, для фонемы $|\bar{o}|$ мы имеем 1510 употреблений в начальной позиции, что соответствует значению 69,89 после нормализации. Так как 69,89 намного больше, чем $|z_{0,05}| = 1,96$, то можем заключить, что фонема $|\bar{o}|$ имеет очень сильную тенденцию к использованию в начале слова. Для фонемы $|\bar{o}|$ в конце слова мы имеем абсолютное значение 0 и нормализованное значение $-20,8$, то есть очень сильную отрицательную тенденцию. Более того, можно даже говорить о том, что фонема $|\bar{o}|$ не может быть использована в абсолютном конце слова в словоформах татарского языка. В анализируемом тексте не встречаются арабские заимствования, оканчивающиеся на букву $\bar{o}$, типа *табиб* ‘врач’. Впрочем, о нетипичности таких слов, об оглушении $|\bar{o}|$ в речи свидетельствует присоединение к ним глухих вариантов словоизменительных аффиксов:

табибка (врач-DIR) ‘к врачу’,
табибта (врач-LOC) ‘у врача’,
табибтан (врач-ABL) ‘от врача’.

Именно такие варианты словоизменения мы обнаруживаем в Татарском национальном корпусе «Туган тел» [Татарский национальный корпус 2022].

Также очень высокую z -статистику для начала слова имеет фонема $|\kappa|$, а именно 57,91 при 2059 случаях употребления в абсолютном начале слова.

Для фонемы $|\zeta|$ в начальной позиции нормализованное значение равно 1,33, что располагается в интервале $(-1,96, 1,96)$ и свидетельствует о том, что данная величина может быть случайной. Возможно,

низкая частотность данной фонемы в исконно татарских словах восполняется ее частотностью в начальной позиции в арабских заимствованиях, и влияние этих двух противоположных тенденций обеспечивается нейтральное значение z-статистики.

Анализ таблицы показывает, что лишь в достаточно редких случаях z-статистика располагается в интервале $(-1,96, 1,96)$, в частности, это заимствованные фонемы $|y|$ и $|u|$, а также фонема $|ж|$, которая характеризуется очень низкой частотностью.

Таким образом, текстовый материал свидетельствует о том, что фонемы имеют явную тенденцию использоваться в той или иной позиции в тексте, отражая специфику организации фонологической системы языка.

Если выбор морфологических категорий зависит от способа выражения мысли, то фонологическая структура языка тесно связана с морфологическим строем слова [Баскаков 1988: 3]. Поэтому далее дадим небольшой грамматический комментарий к таблице 2.

Самой частотной фонемой является фонема $|a|$; фонемы $|a|$ и $|ə|$ используются в аффиксах множественного числа, в аффиксах падежей с пространственным значением (директив, аблатив, локатив), в аффиксах многих глагольных форм (инфинитива, будущего времени, перфекта, прошедшего времени и причастия на *-ган* и др.).

Фонемы $|ы|$ и $|е|$ используются в формах принадлежности у существительных, в аффиксах генитива и аккузатива, категорического прошедшего времени и др.

Фонемы $|л|$ и $|р|$ используются в аффиксе множественного числа, фонема $|л|$ также часто используется во многих словообразующих аффиксах у слов разных частей речи (*-лы, -лык, -ла, -лан* и пр.).

Фонема $|н|$ используется в аффиксах генитива и аккузатива, в пассивных и рефлексивных формах глаголов и т. п.

Фонема $|к|$ — самая частотная в позиции в начале слова.

Очень большое значение z-статистики для фонемы $|н|$ в финальной позиции, а именно 61,96, можно объяснить тем, что данная фонема используется в конвербах на *-н*, когда словоформа оканчивается аффиксом конверба.

Далее для наиболее частотных 8 фонем (4 гласных и 4 согласных) приведем относительные частоты употреблений в зависимости от позиции в слове (см. таблицу 3).

Таблица 3. Самые частотные фонемы в тексте

Фоне- ма	В начале		В конце		В середине	
	Кол-во	%	Кол-во	%	Кол-во	%
а	881	12,5	1396	19,8	4778	67,7
э	374	8,7	980	22,7	2969	68,7
ы	14	0,31	894	20,1	3550	79,6
е	89	1,9	1152	24,5	3468	73,6
л	70	1,9	183	4,9	3457	93,2
р	64	1,6	990	25,2	2868	73,1
н	205	4,4	1593	33,8	2911	61,8
к	2059	50,6	589	14,5	1423	35,0

То, что татарские фонемы по-разному распределены в абсолютном начале и абсолютном конце словоформ (см. абсолютные и относительные значения в таблицах 2 и 3), свидетельствуют о том, что в татарском языке фонемы, наряду с другими важными факторами (в частности, ударением и гармонией гласных), важны в когнитивных процессах определения границ словоформ. Любая словоформа представляет собой целостность, структуру, а опознавание структур предполагает их различия, а это возможно лишь в случае, если отдельные структуры различаются друг от друга известными признаками, ключ к которым дают именно фонемы [Трубецкой 2000: 41—42]. Любая словоформа существует в языке и «как структура всегда представляет собой нечто большее, нежели сумму его членов (= фонем)», но «в противоположность отдельным фонемам эта целостность не может быть локализована в звуковой оболочке слова» [Трубецкой 2000: 42], именно поэтому те стохастические, на первый взгляд случайные распределения фонем, которые раскрывает квантитативный анализ, возникают под действием внутренних законов, управляющих структурами языка.

Статистические характеристики фонем в процессах маркирования границ словоформ на массиве данных — один из интереснейших вопросов, возникших при исследовании распределения фонем в зависимости от позиции в словоформе, который требует специального исследования.

Анализ двухчленных и трехчленных сочетаний фонем

Системность всех уровней в тюркских языках во многом определяется тем, что «слово в тюркских языках по своему строению изоморфно словосочетанию, а фонологический состав слова изоморфен структуре самого слова» [Баскаков 1988: 3]. Важный аспект при квантитативном анализе фонемного состава языка — сочетаемость фонем. Синтагматическая составляющая фонологической системы языка — комбинаторные свойства фонем — является уникальной для каждого языка. В своей классической работе Н. С. Трубецкой говорит о том, что «...сочетания фонем подчиняются в любом языке своим особым законам или правилам, которые имеют значение только для данного языка и которые необходимо устанавливать для каждого языка отдельно» [Трубецкой 2000: 265]. Языки из всех логически возможных комбинаций фонем допускают лишь относительно небольшое количество двух-, трех- или *n*-членных сочетаний, образующих геометрию характерных признаков [Halle 1995].

В современной компьютерной лингвистике *n*-граммы являются важным статистическим инструментом, а информация о типах и частотности сочетаний фонем необходима для полноты описания фонологической системы языка.

Таблица 4 содержит сведения о 22 наиболее частотных двух- и трехфонемных сочетаниях из повести А. Еники. Далее в работе такие сочетания могут называться биграммами и триграммами соответственно.

Таблица 4. Биграммы и триграммы на материале фонем

Ранг	Биграммы	Количество	Триграммы	Количество
1	ар	1206	лар	453
2	ла	982	ала	325
3	ан	916	енә	308
4	ын	870	ләр	302
5	ен	869	ары	279
6	ка	837	ына	278
7	лә	782	кар	268
8	ән	709	елә	255
9	ал	674	иде	255
10	ер	640	бер	250
11	ул	607	аны	231
12	эр	591	ның	228
13	га	591	ган	212
14	на	586	лән	182
15	ак	582	бул	178
16	бе	582	гән	176
17	ны	575	бел	174
18	да	551	нда	172
19	ки	522	гын	169
20	ел	521	эре	168
21	ры	505	рга	167
22	нә	501	ара	166

Всего в тексте обнаружено 61065 бифонемных сочетаний, при этом количество уникальных биграмм относительно невелико — 685 (из логически возможных 1156 вариантов). Из них 574 бифонемные последовательности встречаются 3 и более раз. 50 наиболее частотных биграмм составляют свыше 43% всех бифонемных сочетаний.

Согласно таблице 4, большая часть высокочастотных биграмм имеет вид гласный + сонорный согласный или сонорный согласный + гласный. Между тем в общей таблице биграмм их удельный вес не настолько большой: биграммы со структурой гласный + сонорный согласный составляют 22,7 процентов, биграммы со структурой сонорный согласный + гласный — 16,3% от общего количества биграмм. Такое неравномерное распределение высокочастотных биграмм с сонорными можно объяснить тем, что сонорные согласные очень частотны в составе словоизменятельных и регулярных словообразующих аффиксов.

В таблице 5 приведены 15 наиболее частотных биграмм, встречающихся в начале слова, то есть это фонемные сочетания, которыми начинаются корни (основы) слов.

Таблица 5. Биграммы, частотные в начале слова

Ранг	Биграммы	Количество
1	ка	588
2	бе	465
3	ки	435
4	ба	346
5	бу	315
6	ја	272
7	кү	251
8	кы	239
9	би	233
10	да	219
11	ал	217
12	ке	212
13	шу	210
14	дә	208
15	ди	206
16	та	199
17	ул	189
18	ид	188
19	то	180

Приведем также краткий лингвистический комментарий к сочетаниям из трех фонем в анализируемом тексте. Выделено всего 42 334 трехфонемных сочетания, из них уникальных — 3038 (при логически возможных 39 304). 1902 триграммы встречаются 3 и более раз. 50 наиболее частотных триграмм составляют свыше 20% от их общего числа.

Интересно, что самое частотное сочетание из трех фонем совпадает с основным — заднерядным — алломорфом множественного числа *-лар* (триграмма, совпадающая с переднерядным вариантом этого аффикса *-лар*, имеет ранг 4). Поскольку, как отмечалось выше, фонемы *|б|* и *|к|* имеют очень сильную тенденцию употребляться в абсолютном начале слова, можно предполагать, что частотные трехфонемные сочетания *кар*, *бер*, *бул*, *бел* в большинстве случаев соответствуют основам слов, а не являются частями аффиксальных цепочек. Это отчасти подтверждается данными из таблицы 6, представляющей 12 самых частотных словоформ из текста.

Таблица 6. Частотные словоформы в тексте А. Еники

Ранг	Словоформа	Перевод	Количество	Длина
1	иде	служеб. гл.	165	3
2	бер	‘один’	124	3
3	белән	‘с’	122	5
4	бик	‘очень’	119	3
5	генә	‘только, лишь’	88	4
6	һәм	‘и’	85	3
7	аның	‘его/ее’	81	4
8	гына	‘только, лишь’	77	4
9	дип	цитатив	71	3
10	инде	‘уже, уж, же’	67	4
11	карчык	‘старушка’	65	6
12	кына	‘только, лишь’	58	4

Таким образом, сочетания фонем, представленные в татарском тексте, весьма разнообразны как по своей структуре, так и по возможности их реализации в разных позициях в пределах конкретных словоформ.

Вопрос о том, какие количественные характеристики фонем обусловлены языковой системой, а какие — выбором индивидуального текста, его стилевой принадлежностью и пр., требует дальнейшего изучения на специально подготовленных выборках текстов.

Заключение

Для полноты картины, изображающей фонологическую систему татарского языка, наряду с качественным описанием фонем, представленным в татарских грамматиках, требуется их количественное описание на достаточно больших объемах лингвистических данных.

Данное исследование строилось на графическом текстовом материале, приведенном (с рядом оговорок, отмеченных выше) к фонологически релевантной форме 1 символ — 1 фонема. Несмотря на то, что использованный подход к фонологической транскрипции не дает исчерпывающего представления о фонемах татарского языка, он, тем не менее, позволяет получить вполне адекватные сведения о распределении и комбинаторных свойствах татарских фонем.

Проведенный количественный анализ фонемного состава татарского текста на материале повести А. Еники «Невысказанное завещание» подтвердил тезис о неравномерном распределении фонем в зависимости от их положения в словоформе — в начале, середине или конце слова. Для оценки смещения частот фонем были использованы нормализованные значения (z -статистики) и квантили нормального распределения.

Было показано, что татарские фонемы очень неравномерно распределены в зависимости от позиции в слове: для нормализованных величин, вычисленных при помощи z -теста, наблюдаются как очень большие, так и очень маленькие значения z -статистики, при этом лишь в отдельных случаях нормализованные значения лежат в интервале $(-1,96, 1,96)$, который требуется законами нормального распределения. Величина отклонения реальных z -статистик от данного интервала свидетельствует о степени неслучайности, о характере упорядоченности употребления фонем в языке.

Значения z находятся в интервале $(-1,96, 1,96)$ лишь у заимствованных фонем $|\text{ц}|$ и $|\text{ш}|$, а также у очень низкочастотной фонемы $|\text{ж}|$. Таким образом, возможные случайные распределения затрагивают лишь очень небольшое количество низкочастотных фонем, что не столько ломает глубоко упорядоченный строй татарской фонологической системы, сколько свидетельствует о ее относительной гибкости.

Правила сочетания фонем, так же, как и их инвентарь, являются уникальными для каждого языка. Анализ двух- и трехфонемных сочетаний показал, что сочетания с высоким рангом соответствуют частям аффиксов или основ слов. Таким образом, распределения фонем и фонемных сочетаний во многом определяются агглютинативным строем татарского языка.

Отметим, что данное исследование и его выводы во многом носят предварительный характер. Вопросы о статистических различиях в исконно тюркских и заимствованных словах (как из восточных, так и из западных языков), а также варьирование, обусловленное выбором текста, его стилистикой, тематикой и т. п., требуют дальнейшего изучения. В перспективе было бы также целесообразно провести аналогичное исследование на данных звукового корпуса.

Сокращения

ABL — аблатив

DIR — директив

LOC — локатив

Литература

Байчура 1953 — У. Ш. Байчура. Гласные татарского языка в свете экспериментальных данных: автореф. дис. ... канд. филол. наук. М.; Л., 1953.

Баскаков 1988 — Н. А. Баскаков. Историко-типологическая фонология тюркских языков / Отв. ред. Э. Р. Тенишев. М., 1988.

Богородицкий 1953 — В. А. Богородицкий. Введение в татарское языкознание в связи с другими тюркскими языками. Казань, 1953.

Гришина 2005 — Е. А. Гришина. Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003—2005. М., 2005. С. 94—110.

Гришина, Савчук 2009 — Е. А. Гришина, С. О. Савчук. Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб., 2009. С. 129—149.

Мороз 2019 — Г. А. Мороз. Слоговая структура адыгейского языка: от данных к обобщениям // Вопросы языкознания. 2019, 2. С. 82—95.

Смирнова, Чистиков 2011 — Н. С. Смирнова, П. Г. Чистиков. Программа анализа фонетических статистик в текстах на русском языке и ее использование для решения прикладных задач в области речевых технологий //

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25—29 мая 2011 г.). Вып. 10. М., 2011. С. 632—643.

Татарская грамматика 1993 — Татарская грамматика / Гл. ред. М. З. Закиев. Т. 1. Казань, 1993.

Татар грамматикасы 2015 — Татар грамматикасы / Ред. Ф. М. Хисамова. Т. 1. Казань, 2015.

Татарский национальный корпус 2022 — Татарский национальный корпус «Туган тел» // <http://tugantel.tatar/?lang=ru>, 2022.

Трубецкой 2000 — *Н. С. Трубецкой*. Основы фонологии / Пер. с нем. А. А. Холодовича; под ред. С. Д. Кацнельсона. М., 2000.

Щербак 1970 — *А. М. Щербак*. Сравнительная фонетика тюркских языков. Л., 1970.

Altmann 2002 — *G. Altmann*. Zipfian linguistics // *Glottometrics*. 2002, 3. P. 19—26.

Altmann 2013 — *G. Altmann*. Aspects of Word Length // *Issues in Quantitative Linguistics* / Eds. R. Köhler, G. Altmann. Vol. 3. Lüdenscheid, 2013. P. 23—38.

Altmann 1980 — *G. Altmann*. Prolegomena to Menzerath's Law // *Glottometrika*. 1980, 2. P. 1—10.

Altmann 2005 — *G. Altmann*. Phonic Word Structure // *Quantitative Linguistik. Ein internationales Handbuch* / Eds. R. Köhler, G. Altmann, R. Piotrowski. Berlin; New York, 2005. S. 191—208.

Berta 2015 — *A. Berta*. Tatar and Bashkir // Eds. L. Johanson, É. Á. C. Johanson. *The Turkic Languages*. London; New York, 2015. P. 283—300.

Best 2008 — *K.-H. Best*. Word Length in Persian // *Glottometrics*. 2008, 16. P. 27—30.

Best 2011 — *K.-H. Best*. Word Length Distribution in French // *Glottometrics*. 2011, 22. P. 57—61.

Bogdanova-Beglarian et al. 2017 — *N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko*. Linguistic Features and Sociolinguistic Variability in Everyday Spoken Russian // *International Conference on Speech and Computer*. Springer, Cham, 2017. P. 503—511.

Bogdanova-Beglarian et al. 2018 — *N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko, E. Baeva*. Towards a Description of Pragmatic Markers in Russian Everyday Speech // *International Conference on Speech and Computer*. Springer, Cham, 2018. P. 42—48.

Brown 2017 — *S. Brown*. The Distribution of Phoneme Inventory and Language Evolution // *CAES*. 2017. Vol. 3.1. P. 22—34.

Burbiel 2018 — *G. Burbiel*. Tatar grammar. A Grammar of the Contemporary Tatar Literary Language. Stockholm; Moscow, 2018.

Chomsky, Halle 1968 — *N. Chomsky, M. Halle*. The Sound Pattern of English. Cambridge, Massachusetts, 1968.

Coloma 2014 — *G. Coloma*. Towards a Synergetic Statistical Model of Language Phonology // *Journal of Quantitative Linguistics*. 2014. P. 100—122.

Galieva, Vavilova 2021 — *A. Galieva, Z. Vavilova*. Initial and Final Syllables in Tatar: from Phonotactics to Morphology // *Glottometrics*. 2021, 50. P. 57—75.

Halle 1995 — *M. Halle*. Feature Geometry and Feature Spreading // *Linguistic Inquiry*. 1995, 26 (1). P. 1—46.

Jakobson, Halle 1956 — *R. Jakobson, M. Halle*. Fundamentals of Language. S.-Gravenhage, 1956.

Jakobson, Waugh 1979 — *R. Jakobson, L. Waugh*. The Sound Shape of Language. Brighton, 1979.

Kessler 2015 — *B. Kessler*. Computational and Quantitative Approaches to Historical phonology // *Oxford handbook of historical phonology* / Eds. P. Honeybone, J. Salmons. Oxford, 2015. P. 171—188.

Kromer 2002 — *V. Kromer*. Zipf's law and its modification possibilities // *Glottometrics*. 2002, 5. P. 1—13.

Kromer 2001 — *V. Kromer*. Word length model based on the one-displaced Poisson-uniform distribution // *Glottometrics*. 2001, 1. P. 87—96.

Mandelbrot 1965 — *B. B. Mandelbrot*. Information Theory and Psycholinguistics // *Scientific Psychology* / Eds. B. Wolman, E. Nagel. New York, 1965. P. 550—562.

Moran, McCloy 2019 — *S. Moran, D. McCloy*. PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History, 2019 // <http://phoible.org>, 2019.

Popescu et al. 2015 — *I. Popescu, M. Lupea, D. Tatar, G. Altmann*. Quantitative Analysis of Poetic Texts // *Quantitative Linguistics*. Vol. 67. Berlin, 2015.

Poppe 1963 — *N. Poppe*. Tatar manual: descriptive grammar and texts with a Tatar-English glossary. Bloomington: Indiana University; The Hague: Mouton and Co., 1963.

Radojičić et al. 2019 — *M. Radojičić, B. Lazić, S. Kaplar, R. Stanković, I. Obradović, J. Mačutek, L. Leššová*. Frequency and Length of Syllables in Serbian // *Glottometrics*. 2019, 45. P. 114—123.

Thomson 2016 — *R. I. Thomson*. Does training to perceive L2 English vowels in one phonetic context transfer to other phonetic contexts? // *Canadian Acoustics*. 2016, 44 (3). P. 198—199.

Turunen, Lipping 2008 — *J. Turunen, T. Lipping*. Phoneme analysis based on quantitative and qualitative entropy measurement // *Computer Speech & Language*. 2008, 22 (4). P. 313—329.

Zörnig et al. 2019 — P. Zörnig, K. Stachowski, A. Ráková, Y. Qu, M. Mistecký, K. Ma, M. Lupea, E. Kelih, V. Gröller, H. Gnatchuk, A. Galieva, S. Andreev, G. Altmann. Quantitative Insights into Syllabic Structures // *Studies in Quantitative Linguistics* 30. Lüdenscheld, 2019.

References

- Altmann 2002 — G. Altmann. Zipfian linguistics In: *Glottometrics*. 2002. 3. P. 19—26.
- Altmann 2013 — G. Altmann. Aspects of Word Length. In: *Issues in Quantitative Linguistics* / Eds. R. Köhler, G. Altmann. Vol. 3. Lüdenscheld, 2013. P. 23—38.
- Altmann 1980 — G. Altmann. Prolegomena to Menzerath's Law. In: *Glottometrika*. 1980. 2. P. 1—10.
- Altmann 2005 — G. Altmann. Phonic Word Structure. In: *Quantitative Linguistik. Ein internationales Handbuch* / Eds. R. Köhler, G. Altmann, R. Piotrowski. Berlin; New York, 2005. P. 191—208.
- Baychura 1953 — U. Sh. Baychura. Glasnye tatarskogo jazyka v svete eksperimental'nyh dannyh: avtoref. dis. ... kand. filol. nauk. M.; L., 1953. {U. Sh. Baychura. Vowels of the Tatar language in the light of experimental data: abstract of dis. ... cand. philol. sciences. Moscow, Leningrad, 1953.}
- Baskakov 1988 — N. A. Baskakov. Istoriko-tipologicheskaja fonologija tjurkskih jazykov / Ed. E. R. Tenishev. M., 1988. {N. A. Baskakov. Historical and typological phonology of Turkic languages / Ed. E. R. Tenishev. Moscow, 1988.}
- Berta 2015 — A. Berta. Tatar and Bashkir. In: L. Johanson, É. Á. C. Johanson (eds.). *The Turkic Languages*. London; New York, 2015. P. 283—300.
- Best 2008 — K.-H. Best. Word Length in Persian. In: *Glottometrics*. 2008. 16. P. 27—30.
- Best 2011 — K.-H. Best. Word Length Distribution in French. In: *Glottometrics*. 2011. 22. P. 57—61.
- Bogdanova-Beglarian et al. 2017 — N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko. Linguistic Features and Sociolinguistic Variability in Everyday Spoken Russian. In: *International Conference on Speech and Computer*. Springer, Cham, 2017. P. 503—511.
- Bogdanova-Beglarian et al. 2018 — N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko, E. Baeva. Towards a Description of Pragmatic Markers in Russian Everyday Speech. In: *International Conference on Speech and Computer*. Springer, Cham, 2018. P. 42—48.
- Bogorodickij 1953 — V. A. Bogorodickij. Vvedenie v tatarskoe jazykoznanie v svjazi s drugimi tjurkskimi jazykami. Kazan, 1953. {V. A. Bogorodickij. Introduction to Tatar linguistics in connection with other Turkic languages. Kazan, 1953.}
- Brown 2017 — S. Brown. The Distribution of Phoneme Inventory and Language Evolution In: *CAES*. 2017. Vol. 3. 1. P. 22—34.
- Burbiel 2018 — G. Burbiel. Tatar grammar. A Grammar of the Contemporary Tatar Literary Language. Stockholm — Moscow, 2018.
- Chomsky, Halle 1968 — N. Chomsky, M. Halle. *The Sound Pattern of English*. Cambridge, Massachusetts, 1968.
- Coloma 2014 — G. Coloma. Towards a Synergetic Statistical Model of Language Phonology In: *Journal of Quantitative Linguistics*. 2014. P. 100—122.
- Galieva, Vavilova 2021 — A. Galieva, Z. Vavilova. Initial and Final Syllables in Tatar: from Phonotactics to Morphology. In: *Glottometrics*. 2021. 50. P. 57—75.
- Grishina 2005 — E. A. Grishina. Ustnaja rech' v Nacional'nom korpuse russkogo jazyka. In: *Nacional'nyj korpus russkogo jazyka: 2003—2005*. M., 2005. S. 94—110. {E. A. Grishina. Oral speech in the National Corpus of the Russian Language. In: *National Corpus of the Russian Language: 2003—2005*. Moscow, 2005. P. 94—110.}
- Grishina, Savchuk 2009 — E. A. Grishina, S. O. Savchuk. Korpus ustnyh tekstov v NKRJa: sostav i struktura. In: *Nacional'nyj korpus russkogo jazyka: 2006—2008. Novye rezul'taty i perspektivy*. St. Petersburg, 2009. S. 129—149. {E. A. Grishina, S. O. Savchuk. A Corpus of Oral Texts in the NCRL: Composition and Structure. In: *National Corpus of the Russian Language: 2006—2008. New results and prospects*. St. Petersburg, 2009. P. 129—149.}
- Halle 1995 — M. Halle. Feature Geometry and Feature Spreading. In: *Linguistic Inquiry*. 1995. 26 (1). P. 1—46.
- Jakobson, Halle 1956 — R. Jakobson, M. Halle. *Fundamentals of Language*. S.-Gravenhage, 1956.
- Jakobson, Waugh 1979 — R. Jakobson, L. Waugh. *The Sound Shape of Language*. Brighton, 1979.
- Kessler 2015 — B. Kessler. Computational and Quantitative Approaches to Historical phonology. In: *Oxford handbook of historical phonology* / Eds. P. Honeybone, J. Salmons. Oxford, 2015. P. 171—188.
- Kromer 2002 — V. Kromer. Zipf's law and its modification possibilities. In: *Glottometrics*. 2002. 5. P. 1—13.
- Kromer 2001 — V. Kromer. Word length model based on the one-displaced Poisson-uniform distribution. In: *Glottometrics*. 2001. 1. P. 87—96.
- Mandelbrot 1965 — B. B. [Mandelbrot]. *Information Theory and Psycholinguistics*. In: *Scientific Psychology* / Eds. B. Wolman, E. Nagel. New York, 1965. P. 550—562.

Moran, McCloy 2019 — S. Moran, D. McCloy. PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History, 2019. In: <http://phoible.org>, 2019.

Moroz 2019 — G. A. Moroz. Slogovaja struktura adyghejskogo jazyka: ot dannyh k obobshhenijam. In: *Voprosy jazykoznanija*. 2019. 2. S. 82—95. {G. A. Moroz. The syllable structure of the Adyghe language: from data to generalizations. In: *Voprosy jazykoznanija*. 2019. 2. P. 82—95.}

Popescu et al. 2015 — I. Popescu, M. Lupea, D. Tatar, G. Altmann. Quantitative Analysis of Poetic Texts // *Quantitative Linguistics*. Vol. 67. Berlin, 2015.

Poppe 1963 — N. Poppe. Tatar manual: descriptive grammar and texts with a Tatar-English glossary. Bloomington: Indiana University; The Hague: Mouton and Co., 1963.

Radojičić et al. 2019 — M. Radojičić, B. Lazić, S. Kaplar, R. Stanković, I. Obradović, J. Mačutek, L. Leššová. Frequency and Length of Syllables in Serbian. In: *Glottometrics*. 2019. 45. P. 114—123.

Shcherbak 1970 — A. M. Shcherbak. *Sravnitel'naja fonetika tjurkskih jazykov*. L.: Nauka, 1970. {A. M. Shcherbak. Comparative phonetics of Turkic languages. Leningrad: Nauka, 1970.}

Smirnova N., Chistikov 2011 — N. Smirnova, P. Chistikov. Programma analiza foneticheskikh statistik v tekstah na russkom jazyke i eje ispol'zovanie dlja reshenija prikladnyh zadachah v oblasti rechevyh tehnologij. In: *Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog»* (Bekasovo, 25—29 maja 2011 g.). Vol. 10. M., 2011. S. 632—643. {N. Smirnova, P. Chistikov. A program for the analysis of phonetic statistics in texts in Russian and its use for applied solutions in the field of speech technologies. In: *Computational Linguistics and Intelligent Technologies: Based on the materials of the “Dialogue” annual International Conference* (Bekasovo, May 25—29, 2011). Vol. 10. Moscow, 2011. P. 632—643.}

Tatarskaja grammatika 1993 — Tatarskaja grammatika / Ed. M. Z. Zakiev. Vol. 1. Kazan, 1993. {Tatar grammar / Ed. M. Z. Zakiev. Vol. 1. Kazan, 1993.}

Tatar grammatikası 2015 — Tatar grammatikası / Ed. F. M. Khisamova. Vol. 1. Kazan, 2015. {Tatar grammar / Ed. F. M. Khisamova. Vol. 1. Kazan, 2015.}

Tatar National Corpus 2022 — “Tugan Tel” Tatar National Corpus. In: <http://tugantel.tatar/?lang=en>, 2022.

Trubeckoj 2000 — N. S. Trubeckoj. *Osnovy fonologii* / Ed. S. D. Kacnel'son. M., 2000. {N. S. Trubeckoj. Fundamentals of phonology / Ed. S. D. Kacnel'son. Moscow, 2000.}

Thomson 2016 — R. I. Thomson. Does training to perceive L2 English vowels in one phonetic context transfer to other phonetic contexts? In: *Canadian Acoustics*. 2016. 44 (3). P. 198—199.

Turunen, Lipping 2008 — J. Turunen, T. Lipping. Phoneme analysis based on quantitative and qualitative entropy measurement. In: *Computer Speech & Language*. 2008. 22 (4). P. 313—329.

Zörnig et al. 2019 — P. Zörnig, K. Stachowski, A. Ráková, Y. Qu, I. Místecký, K. Ma, M. Lupea, E. Kelih, V. Gröller, H. Gnatchuk, A. Galieva, S. Andreev, G. Altmann. Quantitative Insights into Syllabic Structures. *Studies in Quantitative Linguistics* 30. Lüdenscheid, 2019.