Glottometrics 50, 2021, 57-75.

Initial and Final Syllables in Tatar: from Phonotactics to Morphology

Alfiya Galieva[1]
Zhanna Vavilova[2]

Abstract

The paper proposes a methodology for analyzing the syllabic structure of Tatar words using fiction text data. Syllable construction rules are unique for each language as they are determined by the laws that govern its specific internal structure. However, the issue of the syllable finds a rather superficial description in Tatar grammars. Thus, possible correlations of the syllable structure with morphological features of the language will be examined in this paper. We analyze the distribution of syllable types in Tatar texts and represent their ranked frequencies and theoretical values fitted by means of the Zipf-Mandelbrot distribution. The main part of the study is devoted to inquiry into the structure of initial and final syllables. We proceed from the hypothesis that distributions of syllable structures in word-initial and word-final positions should be marked by statistically important differences due to discriminative structural features of stems and affixal chains. The study is based on a selection of obstruent and sonorant consonants. To evaluate statistical significance of these differences, the well-known χ^2 test is applied.

Keywords: *syllable, syllable structure, the Tatar language, phonotactics and morphology, quantitative linguistics.*

1. Introduction

Discovering statistically significant connections and consistent patterns between different levels of a language is a task that is successfully solved by means of quantitative linguistics. A discovery of such dependencies can provide new information about the internal structure of the language and the laws that govern it. Linguistic phenomena, despite certain deviations and diversity in their behavior, are characterized by a well-defined regularity and a stable relative frequency. Texts, as manifestations of languages, consist of a large number of elements of different nature whose connections are complex, being influenced by random factors. So text data provide us with information on language regularities within certain variances.

Syllables constitute the most important level between the meaningless (phonemes) and the meaningful (morphemes and words) language units. Many languages of the world have a fixed syllabic structure: possible combinations of phonemes in a syllable are rigidly determined. Relations between the phonotactic organization of words and the morphology of the language is an issue yet poorly described in linguistics. Thus, developing a methodology of such research is a topic of current interest.

The main task of this paper is to propose a method for discovering possible relationships between syllable patterns and morphological features of Tatar, a language with a rich agglutinative morphology. We compare the structure of the initial syllables of polysyllabic words (which are stems or parts of stems) and that of the final syllables (which are mainly affixes or parts of affixal chains) and conduct special tests to determine statistically significant

[1] Kazan Federal University, Kazan, Russian Federation, amgalieva@gmail.com. http://orcid.org/0000-0003-2915-4946.
[2] Kazan State Power Engineering University, Kazan, Russian Federation, zhannavavilova@mail.ru. http://orcid.org/0000-0002-0247-8257.

differences between them. Classical and modern texts of Tatar literature serve as an empirical source for the study.

The body of the paper is organized as follows: Section 2 covers research background. Section 3 contains basic information on Tatar as demanded by the study goals. Section 4 outlines the main stages of data preparation. Section 5 represents the quantitative data on syllable structures in the analyzed texts and the results of the χ^2 test with evaluation of statistical significance of differences between structural features of the initial and final syllables in polysyllabic words; Section 6 concludes and outlines the prospects of future work.

Tatar words, except in cases when it is necessary to represent their original graphical form, are introduced in the extended Latin transcription.

2. Research background

In recent decades, a large number of syllable studies of different languages of various types and structures have appeared, providing researchers with multiple perspectives on the topic. A book called *The Notion of Syllable across History, Theories and Analysis* edited by D. Russo (2015) introduces investigations into the syllable from four points of view: historical, descriptive, analytical-instrumental, and theoretical. Discussions on the nature and structure of the syllable bring into question both the status of the minimal unit of language and methods of linguistic analysis.

Distinguishing segments within the syllable depends on the theoretical assumptions of researchers and on the language type, so it differs in various approaches. Harry van der Hulst and Nancy A. Ritter distinguish onset-rhyme models, mora models and hybrid models (1999: 22-38). The most frequently model subdivides the syllable into three constituents: onset, nucleus, and coda (Haugen, 1956; Davis, 1988); see Figure 1. This last approach will be used to analyze the syllabic structure of Tatar words.

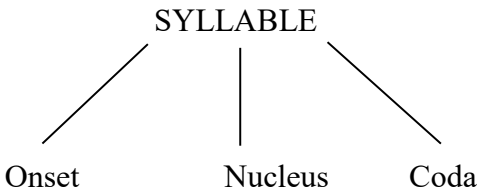

Figure 1. The syllable constituents

The onset is a constituent comprising the syllable-initial consonant or consonant cluster; the nucleus consists of the vowel or syllabic consonant and is considered the peak of the syllable and its obligatory constituent; the coda contains the syllable-final consonant or consonant cluster (Davis, 2006).

Many researchers address the problem of ranking the constraints that determine the syllable structure in a particular language by creating syllable division models within the Optimality Theory (Prince & Smolensky, 1993; Féry & van de Vijver, 2003).

Languages may differ by the degree of distinctness of syllabication; from this perspective, the so-called quantum and wave languages are distinguished. The former has clear-cut syllables with a strictly defined structure, whereas in wave languages, the structures of consonant combinations are vague, with fuzzy syllabic boundaries – so fuzzy that native speakers can mark them in different ways (Kodzasov & Muravyova, 1980). According to this classification, Tatar belongs to quantum languages.

Syllable structure impacts word length (Grzybek, 2007; Antić, Kelih, & Grzybek, 2007), language complexity (Fenk, Fenk-Oczlon, & Fenk, 2006), as well as other linguistic phenomena and has a variety of dimensions (Russo, 2015; Zörnig et al., 2019).

Researchers use different classifications of phonemes when analyzing syllabic structures. Some do not distinguish consonant subclasses (for example, Zörnig et al., 2019). In Russian linguistics, obstruent and sonorant consonants are traditionally opposed (Knyazev, 2006), which is often taken into consideration by researchers of languages of the Russian Federation. For example, Moroz (2019) describes the syllable structure of Adyghe, a Northwestern Caucasian language spoken in Russia and some other countries. Because of unclear syllabification rules in Adyghe, the author uses only word-initial syllable onsets and word-final codas of items taken from an Adyghe-Russian dictionary, thereby disregarding word-medial consonant clusters. It is revealed that the structures of onsets and codas in Adyghe differ, and that nearly all attested consonantal classes can occur in both positions.

An international research team inspired by G. Altmann published a book called *Quantitative Insights into Syllabic Structures* (Zörnig et al., 2019) which generalized data of languages with different morphologies and represented a quantitative analysis of syllable types, syllable length, open and closed syllables, asymmetry of onsets and codas, distances, and syllabic sequences in German, Polish, Slovak, Slovene, Russian, Romani, Chinese, Tatar and some other languages. The book examines a variety of statistical methods applied to the syllable data, and compares theoretical values with the empirical ones provided by languages of different types and origin.

Quantitative studies of linguistic structures, including the types and structure of syllables, based on the evidence of individual languages, raise the question of validity of the existing models (considering a significant number of fluctuations in empirical data, creativity of the text production process, etc.) (Altmann & Gerlach, 2016). Thus, the notion of syllable, seeming intuitively clear, remains at the intersection of discussions, and syllabic phenomena may serve as a good field for discovering alternative solutions for building language models.

As for Turkic languages, where Tatar belongs, there are special tools used to extract syllables. In particular, TASA (Turkish Automatic Spelling Algorithm) was developed for Turkish and was tested over five different corpora (Aşliyan & Günel, 2005). In Tatar linguistics, the concept of syllable remains on its periphery, although modern computer technologies make it easy to develop tools for automatically selecting syllable structures, as well as for their quantitative study. The available Tatar grammars pay very little attention to the problem of syllabication and syllable structure (Zakiev, 1993: 85-87; Khisamova, 2015: 40-41), and the number of special studies is very limited (Galieva, 2020). Therefore, quantitative analysis of syllable structures would be the first and a very important step to building a model of the syllable based on Tatar language data.

3. Overview of Tatar

Tatar, a Turkic language, is spoken in the Volga and Ural regions of Russia and in some regions of Siberia. It is the second most common language of the Russian Federation and is, on a par with Russian, the official language of the Republic of Tatarstan. According to the 2010 Census, the number of Tatar speakers in Russia is 5.31 million people (Vserossijskaya perepis, 2010).

The location of Tatar culture at the intersection of Occidental and Oriental civilizations leads to active language contacts both with the Arab-Muslim and the European cultural areas. A significant part of abstract vocabulary in Tatar is of Oriental (Arabic and Persian) origin, and

many scientific and technical terms come from Europe. Historical contacts with Russian and its current dominant role as the state language of the Russian Federation became a cause of a huge number of words and constructions borrowed and calqued (component-by-component translated) from Russian. Consequently, modern Tatar has a large number of synonymous items of different – Turkic, Russian, European, and Oriental – origin (Galieva, 2018). As a rule, oriental loanwords were borrowed centuries (and even millennia) ago, which resulted in their phonetic assimilation. Unlike those, loanwords of European origin appeared through Russian mediation during the last centuries, maintaining a graphical and phonological shape typical for Russian.

In the 8th – 10th centuries, ancient Turkic peoples used a runic script – the so-called Orkhon-Yenisey script, named after the Orkhon Valley in Mongolia. After Volga Bulgars converted to Islam in 922, thus establishing firm contacts with Arabic and Muslim cultural areas, the ancient runic script was replaced by the Arabic script to be used by the ancestors of modern Tatars for over a millennium. In 1928, the Arabic script was changed to Latin. Since 1938 – 1940, Cyrillic is the official Tatar alphabet which employs all Russian letters and 6 additional ones to designate specific Tatar sounds. Therefore, the total number of letters in present Tatar alphabet is 39. Nevertheless, this script maps pronunciation of Tatar words not consistently enough, allowing for variations in writing and ambiguities in reading.

Modern Tatar has a rich system of phonemes: it includes 9 original vowels and 3 additional ones used in loanwords; the consonant system comprises 25 original and 5 additional consonants used in loanwords. Sonorant consonants in Tatar include glides *j* and *w*, liquids *r* and *l*, and nasals *m, n, ŋ* (Zakiev, 1993; Khisamova, 2015). In original Tatar, there are no affricates, and obstruents are divided into stops and fricatives.

According to Tatar grammars, original Tatar words are constructed from syllables of six types: V, CV, VC, CVC, VSC, CVSC[3]; some other syllable types can be found in loanwords (Zakiev, 1993: 85; Khisamova, 2015: 40).

The most important phonetic feature of Turkic languages is progressive vowel harmony. In Tatar, vowel harmony is a morphonological assimilatory process involving agreement between vowels within a word form. Original Tatar one-root words contain only back vowels *(a, o, u, ı*) or only front vowels *(ä, ö, ü, e, i*). Due to progressive (from-beginning-to-end) direction of vowel harmony, the quality of vowels in affixes and in affixal chains is determined by the quality of vowels in stems; the latter do not alternate, leaving it for alternating affixes (derivational and inflectional ones) to follow the vowel harmony rules.

This is how vowel harmony works for a word containing back vowels:
Bala ‘child’
Bala-lar-da
Child-PL, LOC[4]
‘in children’
This is how vowel harmony works for a word containing front vowels:
Mäktäp ’school’
Mäktäp-lär-dä
School-PL, LOC
‘in schools’

[3] C – obstruent consonant, S – sonorant consonant, V – vowel.
[4] PL – Plural, LOC – Locative.

Some Tatar particles of Turkic origin also obey the vowel harmony law which governs the co-occurrence of vowels within a span of utterance. For example, particles *da* 'too, also' and *gına* 'only, merely', like affixes, have phonetic variants depending on the nature of the preceding word:

Ber ***bala gına*** 'only one child'

Ber ***kön genä*** 'only a day' .

Such items form phonological words according to vowel harmony rules.

Words with vowel harmony violations contain a mixed set of vowels – front and back vowels at the same time. These exceptions are mainly compound words consisting of two or more stems or loanwords (from Arabic, Persian, European languages or Russian), for example:

Kitap 'book', Arabic loanword;

Tarih 'history', Arabic loanword;

Maşina 'machine', Greek loanword.

Tatar is characterized by rich agglutinative morphology. The basic way of word formation and inflection is progressive affixal agglutination when a new unit is built by consecutive addition of regular and clear-cut monosyllabic derivational and inflectional affixes to the stem. The boundaries between the affixes within the word form are distinct and transparent, so that the affixal joint in many cases coincides with the syllabication (Guzev & Burykin, 2007).

4. Data preparation

The preparatory stage of the research included the following main steps:

- selection of linguistic data;
- conversion of written text items into phonological form;
- dividing words into syllables.

We did not use data of Tatar dictionaries for several reasons:

- they contain a large number of loanwords with a phonological structure which is not typical for Tatar (for example, items with complex consonant clusters), and a great number of these words are rarely used in real texts;

- they fix words in basic forms with their typical structure (for example, verbs in Tatar are given in the Infinitive or Verbal Noun forms);

- they do not map inflected forms of words, and those are crucial for analyzing Tatar with its agglutinative morphology.

The objective was to study the distribution of syllables in real use, so we consider textual material to be the best source. Therefore, fiction texts of different genres (poetry and prose by Tatar classical and modern writers) were selected as a source of Tatar language patterns (basic information on this selection is given in Appendix; the titles of the texts are given in transliteration and in translation).

The next stage was bringing the written text to the standard form: 1 letter – 1 sound. It is believed that Tatar writing is generally based on exactly this principle (nevertheless, there are some exceptions). For this purpose, it seems to be relevant to mention here the main features of Tatar spelling.

1. In Tatar, there are two letters (*ь* and *ъ*) that do no denote any sound but determine the pronunciation of adjacent letters.
2. Letters *я* and *ю* denote correspondingly a couple of sounds *ya/ yä* or *yu/ yü* (the choice of *a / ä* and *u / ü* is determined by the vowel structure of the word).

3. *Е* may by pronounced as *ye, yı* or *e* depending on its position in the word and the word vowel structure.
4. *У* and *ү* after *а / ә* are pronounced as *w* sonorant consonant and as *u* or *ü* vowels in any other case.
5. Besides, *в* may be pronounced as *v* in Russian and European loanwords and as *w* in original Tatar and Oriental (Arabic and Persian) loanwords.

So special rules were set to convert Tatar texts into a phonologically relevant form.

Then phonological structure of words was mapped as frames consisting of vowels, sonorant (*l, r, m, n, ŋ, w, j*) and obstruent consonants. The differentiation between sonorant and obstruent consonants fits well for modeling syllables in Turkic languages.

Next, rules of dividing words into syllables were developed, and syllables were mapped basing on the available grammars of the Tatar language (Zakiev, 1993; Khisamova, 2015).

Table 1.

Main stages of word analysis

Original Cyrillic word form	**Phonological mapping of the word**	**Syllable structure of the word**
урман ‘forest, wood’	/urman/	VS-SVS
егет 'young man'	/yeget/	SV-CVC
ямьле ‘nice’	/yämle/	SVS-SV
аулау 'to hunt'	/awlaw/	VS-SVS
юл 'road, way'	/yul/	SVS

At the last stage, the data were statistically processed and the results were visualized[5].

5. Syllable structures in Tatar

5.1. Main syllable patterns

In the framework of our study, it is important to divide consonants into obstruents and sonorants according to the ratio of voice and obstructing airflow. First, let us see how syllables of different structures are distributed in the Tatar texts. Table 2 represents the distribution of syllable patterns in 10 Tatar texts (only 10 most frequent syllable types are presented). The table shows that syllables of simple structure, composed of an initial consonant (obstruent or sonorant one, CV and SV types together) and a vowel, make up about 40-50% of all syllables. Syllables consisting of one consonant onset, a nucleus vowel and one consonant coda account for 38-48%. Relative frequencies of syllable patterns in individual texts may differ significantly. The CV syllable has rank 1 in all the texts processed, in other words, it is the most frequent type. The SV and CVS syllable patterns have rank 2 or 3, depending on the text. The CVS type has rank 2 or 3, and the SVS type has rank 5 in most of the texts. It is noteworthy that the SVS type is absent among the 6 canonical syllable types presented in Tatar grammars; perhaps this is due to the fact that the SVS pattern occurs mainly at the end of words in affixal chains (see Figure 3 where the distribution of types of syllables in word-initial and word-final positions is presented).

[5] All the stages were implemented in R programming language (R Core Team, 2018); besides, tidyverse (Wickham 2017) and stringr (Wickham 2019) packages were used.

Table 2.
Frequencies of syllables of different structures in Tatar texts

Type	Eniki			Tukay, *Şüräle*			Alish		
	Rank	**Obs.**	**Expected**	**Rank**	**Obs.**	**Expected**	**Rank**	**Obs.**	**Expected**
CV	1	1531	1482	1	506	545	1	559	574
SV	2	897	1037	3	268	279	3	313	315
CVS	3	776	729	2	392	390	2	444	428
SVS	4	471	510	4	222	200	4	222	225
CVC	5	417	362	5	207	144	5	180	169
V	6	359	255	7	86	74	6	141	124
SVC	7	243	178	6	103	103	7	107	90
VS	8	129	124	8	66	53	8	79	68
VC	9	96	89	9	41	38	9	79	56
CVSC	10	26	59	11	6	19	10	6	34
	s = 125.96, b = 351.96, χ^2 =225.4			s = 124.73, b = 370.31 χ^2 =86.2			s = 123.45, b = 401.78, χ^2 =82.8		

Type	Amirhan			Tukay, *Käcä belän sarık*			Tukay, *Su anası*		
	Rank	**Obs.**	**Expected**	**Rank**	**Obs.**	**Expected**	**Rank**	**Obs.**	**Expected**
CV	1	421	386	1	417	364	1	246	238
SV	2	216	274	3	157	177	3	115	126
CVS	3	210	195	2	190	252	2	170	173
SVS	4	132	139	5	111	89	5	78	67
CVC	5	113	99	4	127	125	4	87	91
V	7	71	50	7	49	46	6	48	49
SVC	6	89	71	6	76	64	7	43	35
VS	8	29	36	8	37	34	8	37	26
VC	9	23	26	9	31	25	9	19	19
CVSC	10	3	18	10	10	18	12	2	7
	s = 124.82, b = 363.59, χ^2 = 57.9			s = 12.88, b = 33.53 χ^2 =47.3			s =123.59, b = 385.36, χ^2 =23.1		

Type	Amirhan			Tukay, *Käcä belän sarık*			Tukay, *Su anası*		
	Rank	**Obs.**	**Expected**	**Rank**	**Obs.**	**Expected**	**Rank**	**Obs.**	**Expected**
CV	1	421	386	1	417	364	1	246	238
SV	2	216	274	3	157	177	3	115	126
CVS	3	210	195	2	190	252	2	170	173
SVS	4	132	139	5	111	89	5	78	67
CVC	5	113	99	4	127	125	4	87	91
V	7	71	50	7	49	46	6	48	49
SVC	6	89	71	6	76	64	7	43	35
VS	8	29	36	8	37	34	8	37	26
VC	9	23	26	9	31	25	9	19	19
CVSC	10	3	18	10	10	18	12	2	7
	s = 124.82, b = 363.59, χ^2 = 57.9			s = 12.88, b = 33.53 χ^2 =47.3			s =123.59, b = 385.36, χ^2 =23.1		

Type	Zulfat		
	Rank	**Obs.**	**Expected**
CV	1	95	99
SV	3	54	54
CVS	2	73	73
SVS	4	42	43
CVC	5	35	29
V	7	19	16
SVC	6	23	22
VS	9	6	9
VC	8	11	12
CVSC	10	1	6
	s =123.15, b = 399.17, χ^2 =7.5		

To compare, Table 2 also puts forward theoretical values calculated on the basis of the Zipf-Mandelbrot distribution which has the following probability mass:

$$(1)\ p(x) = \frac{(x+b)^{-s}}{\sum_{i=1}^{N}(i+b)^{-s}}$$

where $x = 1, 2, \ldots, N$. S, $b > 0$ are shape parameters, x is the rank of the data, and N is the number of ranks. B. Mandelbrot put forward this distribution to estimate word frequencies (Mandelbrot, 1965); the Zipf-Mandelbrot distribution is often used for modeling syllable frequencies (see, for example, Radojičić et al., 2019). In Table 2, we list the computed parameter values and the measure χ^2 for the goodness of fit (in the table, 10 most frequent syllable types are presented and the fit applied only to them). According to our data, *s* parameter in 8 texts lies in the interval from 123 to 126, and *b* parameter lies in the interval from 350 to 402. Two texts

(the tale *Käcä belän sarık* by Tukay with parameters s = 12.88 and b = 33.53 and the text by Gilman with parameters s = 55.27 and b = 156.04) appear to be outliers.

As was noted above, Tatar grammars represent 6 canonical types of syllables and mention that other types can be found in loanwords (Zakiev, 1993: 85; Khisamova, 2015: 40). We found 22 different syllable types in the examined texts, 7 of which are quite frequent and make up at least 5% each. Syllable patterns with rank lower than 10 characterize rather random features of the texts. Syllable types missing in grammar books come from loanwords; besides, they can be found in morpheme junctions in original Tatar word forms, for example the SVS type in the example below:

barmıym (CVS-SVSS)
go-NEG, PRES, 1SG[6]
'I do not go'.

The distribution of units with different ranks and frequencies of syllable types found in the text by Eniki is presented in the chart (Figure 2). Theoretical values are fitted by means of the Zipf-Mandelbrot distribution.

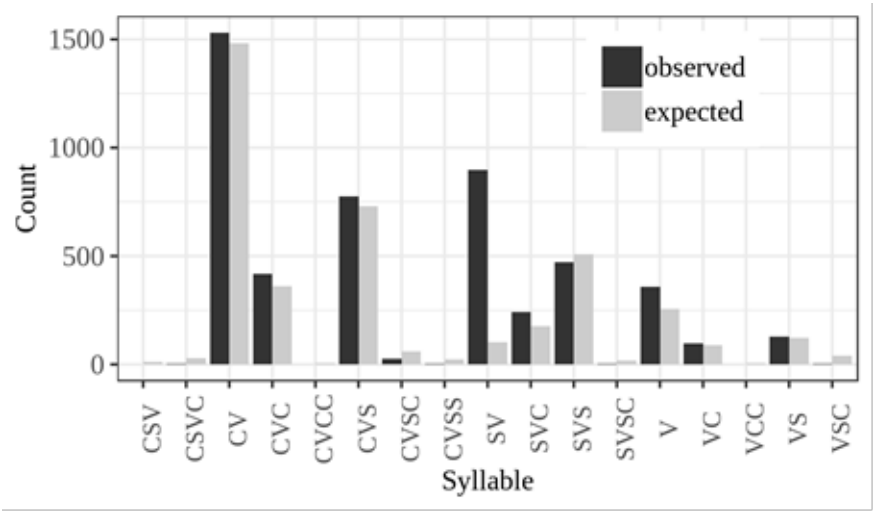

Figure 2. Distribution of syllables of different structures in the text by Eniki

Data represented in Table 2 evidence that complex consonant clusters in syllable onsets and codas occur relatively infrequently. It should be noted that, for example, SC cluster (not syllable type) by itself is quite frequent in Tatar words; however, its elements, when inflected, are often distributed over different syllables. See examples below:

kayt (CVSC) 'return (Imperative)' – *kayta* (CVS-CV) 'he / she returns', kaytaçak (CVS-CV-CVC) 'he / she will return';

kart (CVSC) 'an old man' – *kartı* (CVS-VC) 'his / her old man', *kartım* (CVS-CVS) 'my old man'.

As a result, syllables with consonant SC cluster are relatively rare in our data. So word inflection in many cases simplifies syllabication.

Evidently, new text data, especially texts with numerous loanwords, will provide new patterns of syllable structures with more complex onsets and codas.

[6] NEG – Negative, PRES – Present, 1SG, 1st person, Singular.

5.2. Structure of the initial and final syllables of words

Analysis of the structure of syllables in word-initial and word-final positions interests us because it allows determining how differences in stems and affixes correlate to differences between syllable structures. In Tatar polysyllabic words, initial syllables are stems or parts of stems and final syllables are usually affixes or fragments of affixal chains. Thus, we proceed from the hypothesis that the distribution of syllable structures at the beginning and at the end of word forms should have statistically significant differences; this could be so due to different phonological arrangement of stems and chains of affixes.

The samples included words consisting of two or more syllables taken from 10 Tatar texts; the data was processed separately for each text. For example, in the text by Eniki we detected 1782 words having more than one syllable. So the samples of initial and final syllables comprised 1782 syllables, and the syllables in the middle of words were not considered. 14 types of syllables from total 18 types found in the text by Eniki (taking into account monosyllables and syllables in the middle of the word) are represented in the initial and final positions. Figure 3 represents a number of syllables of different types at the beginning and at the end of word forms in this text. The data indicate that the final syllables tend to fall into a fewer number of types: mainly types CV (474 times), SV (391 times), SVS (296 times), CVS (313 times), SVC (148 times), and CVC (148 times) are represented. Besides, final syllables rarely begin with a vowel (V pattern is not represented in the sample at all, VC pattern is encountered in 3 cases only and VS type is found in 6 cases). Another important feature is that final syllables tend to have a sonorant onset.

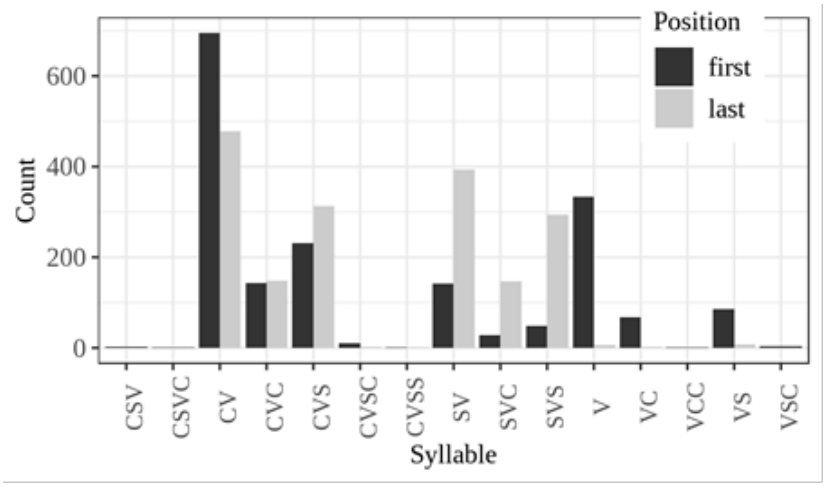

Figure 3. Initial and final syllables defined in the text by Eniki

Initial syllables are characterized by greater diversity, while the CV type dominates, occurring 683 times (26% of the entire sample).

The data presented in Figure 4 also support our assumption that the syllables at the beginning and at the end of words differ in quantitative and qualitative features.

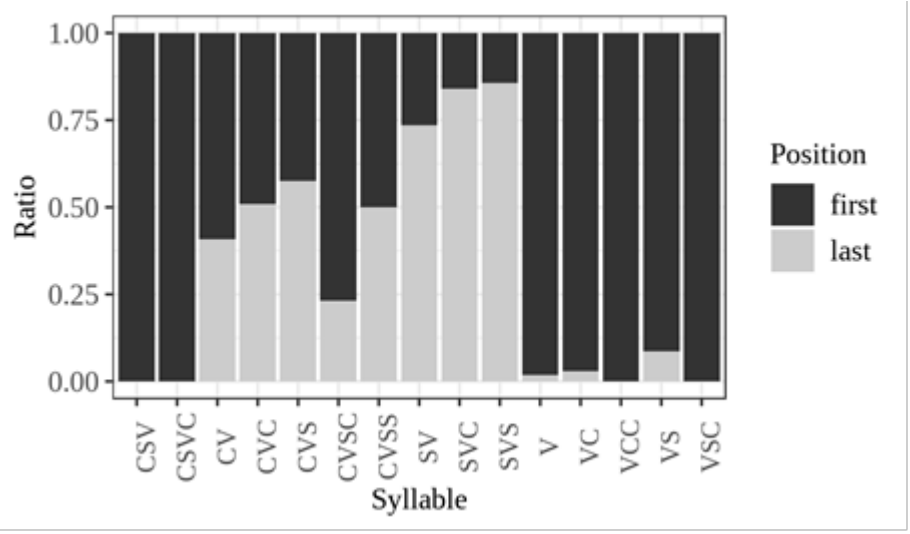

Figure 4. The ratio of initial and final syllables for each type in the text by Eniki

Of frequently occurring syllables, only for the CVC pattern a distribution close to 50% is observed: it occurs 143 times at the beginning of words and 148 times at the end. Although the CVSS type has a 50% distribution, it is characterized by a very low frequency (it occurs once at the beginning and once at the end of the word).

5.3. Initial and final syllables: χ^2 test results

With the obtained data on building initial and final syllables of Tatar words, we can compare the arrangement of onsets and codas in both positions. Now we can ask ourselves whether the distribution of syllables with onsets and codas in initial and final syllables is random. An answer to this question can be provided by applying the χ^2 test, which would allow us to evaluate the statistical significance of differences between nominative variables in a contingency table. In particular, χ^2 criterion of Pearson is a nonparametric method that allows for assessing significance of differences between the observed number of qualitative characteristics of the sample falling into each category and the theoretical amount that can be expected in the studied groups if the null hypothesis is true. The null hypothesis of the χ^2 test is that there is no relationship between columns and rows in the contingency table: the event "an observation in row *i*" is independent of the event "that same observation is in colon *j*" for all *i* and *j* (Conover, 1999: 205). So as far as our data is concerned, the null hypothesis may be formulated as "The proportions of onsets in initial and final syllables are independent". We used the Yates's corrected version of Pearson's χ^2 statistics (Yates, 1934):

$$(2)\ x^2_{Yates} = \sum_{i=1}^{N} \frac{\left(\left|O_i - E_i\right| - 0.5\right)^2}{E_i}$$

where: O_i is observed frequency, E_i is expected frequency, asserted by the null hypothesis, N is the number of distinct events. The generic formula for computing the expected frequency in row *i* and column *j* is given below:

$$(3)\ E_{ij} = \frac{S_i S_j}{N}$$

where s_i is the marginal frequency of row *i,* s_j is the marginal frequency of column *j* and *N* is the total number of observations.

We performed the χ2 test twice, separately for the onsets and the codas. Table 3 presents the results for the onsets.

Table 3.

χ2 test for onsets results

A. Eniki, *Äytelmägän wasıyät*					
Have onsets	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Yes	1299	1530.5	1762	1530.5	3061
No	483	251.5	20	251.5	503
total	1782		1782		3564
	$\chi^2 = 494.07$, df = 1, p-value < 0.0001				

G. Tukay, *Şüräle*					
	Initial syllables		**Final syllables**		**Row total**
Have onsets	**Observed values**	**Expected values**	**Observed values**	**Expected values**	
Yes	512	591	670	591	1182
No	164	85	6	85	170
total	676		676		1352
	$\chi^2 = 165.85$, df = 1, p-value < 0.0001				

G. Tukay, *Şüräle*					
	Initial syllables		**Final syllables**		**Row total**
Have onsets	**Observed values**	**Expected values**	**Observed values**	**Expected values**	
Yes	512	591	670	591	1182
No	164	85	6	85	170
total	676		676		1352
	$\chi^2 = 165.85$, df = 1, p-value < 0.0001				

A. Alish, *Sertotmas ürdäk*					
Have onsets	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Yes	523	625	727	625	1250
No	209	107	5	107	214
total	732		732		1464
	$\chi^2 = 165.85$, df = 1, p-value < 0.0001				

G. Tukay, *Käcä belän sarık äkiyäte*					
Have onsets	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Yes	391	438	485	438	876
No	95	48	1	48	96
total	486		486		972
	$\chi^2 = 99.967$, df = 1, p-value < 0.0001				

G. Tukay, *Su anası*					
Have onsets	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Yes	223	263.5	304	263.5	527
No	82	41.5	1	41.5	83
total	305		305		610
	$\chi^2 = 89.253$, df = 1, p-value < 0.0001				

F. Amirkhan, *Häyät*					
Have onsets	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Yes	359	408	457	408	816
No	100	51	2	51	102
total	459		459		918
	$\chi^2 = 103.78$, df = 1, p-value < 0.0001				

G. Ibragimov, *Kızıl çäçäklär*					
Have onsets	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Yes	290	334.5	379	334.5	669
No	92	47.5	3	47.5	95
total	382		382		764
	$\chi^2 = 93.091$, df = 1, p-value < 0.0001				

G. Gilman, *Oçraşu*					
Have onsets	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Yes	630	713	796	713	1426
No	188	105	22	105	210
total	818		818		1636
	$\chi^2 = 148.73$, df = 1, p-value < 0.0001				

Suleyman, *Dürt mizgel*					
Have onsets	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Yes	225	254	283	254	508
No	67	38	9	38	76
total	292		292		584
	$\chi^2 = 49.146$, df = 1, p-value = 0.0001				

Zulfat, *Söyembikäneŋ huşlaşu dogası*					
Have onsets	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Yes	107	124	141	124	248
No	34	17	0	17	34
total	141		141		281
	$\chi^2 = 36.421$, df = 1, p-value = 0.0001				

An appropriate graphic way for visualizing data from two or more qualitative variables is a mosaic chart. Figure 5 demonstrates that syllables with onsets are strongly overrepresented in the final position and underrepresented in the initial position, whereas syllables with no onsets are strongly underrepresented in the final position. The colour of shading corresponds to the sign of standardized residuals, and the intensity of shading shows relative importance of the differences. The data are represented for the text by Eniki.

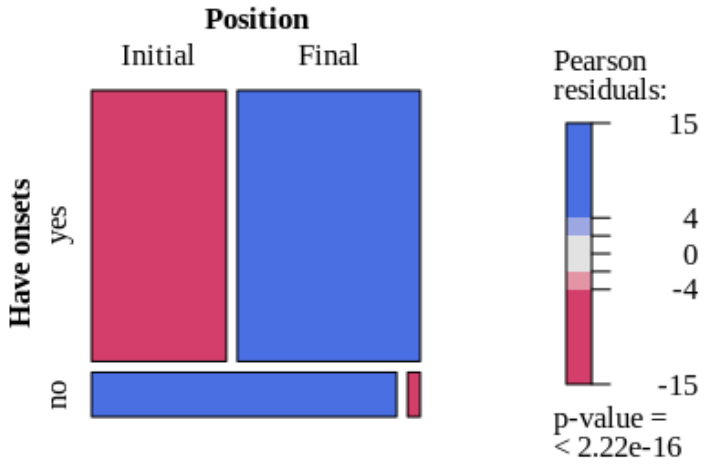

Figure 5. Syllables with and without onsets depending on syllable position

Next, we carried out the χ^2 test for open and closed syllables and found out that the differences between the initial and final syllables are not statistically significant (p-value > 0.05). The result for *Şüräle* by Tukay is presented in Table 4.

Table 4.

χ^2 test results for syllables with coda

Syllable type	**Initial syllables**		**Final syllables**		**Row total**
	Observed values	**Expected values**	**Observed values**	**Expected values**	
Open	326	313	300	313	626
Closed	349	362	375	362	724
Column total	675		675		1350
	$\chi^2 = 1.8617$, df = 1, p-value = 0.1724				

It should be noted that analyzing the first and the last syllables gives but a rough idea of the morphological structure of the Tatar word (the last syllable may be a part of a polysyllabic stem); nevertheless, the main trends can be traced. Thus we examined 10 texts and in all of them found statistically significant differences between patterns of the initial and final syllables.

6. Conclusion

As the aim of the research was to evaluate syllable structures of Tatar word forms in actual use, we analyzed classical and present-day texts of Tatar literature, poetic and prose works or fragments from them, disregarding dictionary data for a large number of loanwords with syllable structures that are atypical for Tatar and for the lack of affixed word forms.

The research design relied upon distinguishing between sonorant and obstruent consonants. It has been found that in Tatar, simple syllable structures prevail (CV, SV, CVS, SVS, CVC, SVC). In many cases, available consonant clusters are broken into joining inflection affixes and fall into two adjacent syllables.

The main task of the study was to propose a way to assess possible correlations between syllable structures and morphology. We analyzed the structure of initial and final syllables of polysyllabic words and compared a number of syllables with and without onsets. The χ^2 test showed that the observed values were statistically significantly different from the expected values.

This study is preliminary in many respects. It is aimed at developing a methodology for studying the structure of the syllable in Tatar in order to create a comprehensive syllable model in the future as well as to disclose possible correlations between phonotactics and morphology. In Tatar, with its rich agglutinative morphology, such correlations should exist and could be quantified. We suppose that further research will be carried out taking into account the sonority scale; analysis of initial and final syllables distinguishing between the types of sonorants (nasals, liquids, and semivowels /w/, /j/), and obstruent consonants (fricatives, stops), it seems, should provide more detailed information about the structure of stems and affixal chains in Tatar.

Acknowledgments: The work is carried out according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

Altmann, E. G., Gerlach, M. (2016). Statistical laws in linguistics. In: *Creativity and Universality in Language*. Springer, 7-26.

Antić, G., Kelih, E., Grzybek, P. (2007). Zero-syllable words in determining word length. Contributions to the science of text and language. In: *Word Length Studies and Related Issues*. Springer, 117 – 156.

Aşliyan, R., Günel, K. (2005). Design and implementation for extracting Turkish syllables and analyzing Turkish syllables. In: *International Symposium on Innovations in Intelligent SysTems and Applications*. INISTA, 170-173.

Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). New York: Wiley.

Davis, S. (1988). *Topics in syllable geometry*. New York: Garland.

Davis, S. (2006). Syllable constituents. In: *The Encyclopedia of Language and Linguistics (2nd ed.)*. Vol. 12. Oxford & New York: Pergamon Press, 326-328.

Fenk, A., Fenk-Oczlon, G., Fenk, L. (2006) Syllable complexity as a function of word complexity. *In The VIII International Conference Cognitive Modeling in Linguistics*. Vol. 1, 324 - 333.

Féry, C. & Vijver van de, R. (eds.) (2003). *The Syllable in Optimality Theory*. Cambridge: Cambridge University Press.

Galieva, A. M. (2018). Synonymy in modern Tatar reflected by the Tatar-Russian Socio-Political Thesaurus. In: Čibej, J. et al. (eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, 585-994.

Galieva, A. M. (2020). Struktura sloga v tatarskom yazyke: ot dannykh k modeli [Syllable structure in Tatar: from data to modeling]. *International Journal of Open Information Technologies* 8 (1), 9-16.

Grzybek, P. (2007) History and methodology of word length studies. In: *Contributions to the Science of Text and Language. Word Length Studies and Related Issues.* Springer, 15-90.

Guzev, V. G., Burykin, A. A. (2007) Obshchie stroevye osobennosti agglyutinativnykh yazykov [General structural peculiarities of agglutinative languages]. In: *Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskikh issledovanii [Papers of Institute of Linguistic Studies, Russian Academy of Sciences]*. Vol. 3-1. Saint-Petersburg: Nestor-istoriya, 109-117.

Haugen, E. (1956). The syllable in linguistic description. In: M. Halle, H. Lunt, & H. McLean (eds.) *For Roman Jakobson*. The Hague: Mouton, 213-221.

Hulst van der, H., Ritter, N. A. (1999). *The Syllable: Views and Facts*. Berlin: Mouton de Gruyter.

Khisamova, F. M. (ed.) (2015). *Tatar grammatikası [Tatar Grammar]*. Vol. 1. Kazan: Institute of Language, Literature and Art.

Knyazev, S. V. (2006). *Struktura foneticheskogo slova v russkom yazyke: sinkhroniya i diakhroniya [Structure of the Phonetic Word in Russian: Synchrony and Diachrony]*. Moscow: Max Press.

Kodzasov, S. V., Muravyova, I. A. (1980). Slog i ritmika slova v alyutorskom yazyke [Syllable and word rhythmics in Alutor]. In: *Publikatsii otdeleniya strukturnoi i prikladnoi lingvistiki MGU. Filologicheskii fakultet [Papers of Department of Structural and*

Applied Linguistics of Moscow State University]. No. 9. Moscow: Lomonosov Moscow State University Press, 103-127.

Mandelbrot, B. B. (1965). Information Theory and Psycholinguistics. In: B. B. Wolman & E.Nagel (eds.) *Scientific Psychology.* New York: Basic Books, 550-562.

Moroz, G. A. (2019). Slogovaya struktura adygeyskogo yazyka: ot dannykh k obobshcheniyam [Adyghe syllable structure: from empirical data to generalizations]. *Voprosy yazykoznaniya [Issues of Linguistics]* 2, 82-95.

Prince, A., Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder. Available from http://roa.rutgers.edu/files/537-0802/537-0802-PRINCE-0-0.PDF

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. Available from https://www.R-project.org/.

Radojičić, M., Lazić, B., Kaplar, S., Stanković, R., Obradović, I., Mačutek, J., Leššová, L. (2019). Frequency and length of syllables in Serbian. *Glottometrics* 45, 114-123.

Russo, D. (2015; ed.). *The Notion of Syllable across History, Theories and Analysis*. Cambridge: Cambridge Scholars Publishing.

Vserossijskaya perepis naseleniya [All-Russian Census] (2010). Available from https://rosstat.gov.ru/free_doc/new_site/perepis2010/croc/perepis_itogi1612.htm

Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. Available from https://CRAN.R-project.org/package=tidyvers

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. Available from https://CRAN.R-project.org/package=stringr

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217-235.

Zakiev, M. Z. (ed.) (1993). *Tatarskaya grammatika [Tatar Grammar]*. Vol. 1. Kazan: Tatar Publishing House.

Zörnig, P., Stachowski, K., Rácová, A., Qu, Y., Místecký, M., Ma, K., Lupea, M., Kelih, E., Gröller, V., Gnatchuk, H., Galieva, A., Andreev, S., Altmann, G. (2019). *Quantitative Insights into Syllabic Structures. Studies in Quantitative Linguistics 30.* Lüdenscheid: RAM-Verlag.

Appendix I

Basic information on the texts processed

No	Author	Title	Genre	Words	Syllables
1	Eniki, Amirkhan	Äytelmägän wasıyät / Unspoken Testament, Chapter 1	novel, prose	2,169	4,962
2	Tukay, Gabdulla	Şüräle / Forest Spirit	fairy tale (verse)	925	1,917
3	Tukay, Gabdulla	Su anası / Aquatic Woman	fairy tale (verse)	419	854
4	Tukay, Gabdulla	Käcä belän sarık äkiyäte/ The tale of the goat and the ram	fairy tale (verse)	579	1,211
5	Amirkhan, Fatikh	Häyät Hayat, Chapter 1	novel, prose	548	1,310
6	Ibrahimov, Galimjan	Kızıl çäçäklär /	novel, prose	444	1,085

		The Red Flowers, Chapter 1			
7	Alish, Abdulla	Sertotmas ürdäk / The Talkative Duck	fairy tale for children, prose	917	2,093
8	Gilman, Galimdzhan	Oçraşu / Встреча	story, prose	1,014	2,351
9	Suleyman	Dürt mizgel / Four moments	poem	355	815
10	Zulfat	Söyembikäneŋ huşlaşu dogası / The farewell prayer of Suyumbike	poem	163	360