Word Length in Tatar: Selecting Relevant Parameters for Modeling

Alfiya Galieva[0000-0003-2915-4946]

Kazan Federal University, Kremlyovskaya St, 18, 420008 Kazan, Russia
amgalieva@gmail.com

Abstract. This paper studies word length in the Tatar language examining data of fiction texts (the sample includes examples of both prose and poetry). Word length is a stochastic phenomenon depending on a great number of factors, including language type, text organization, its addressee, etc.; however, there are internal linguistic laws governing parameters of word length and frequencies of words, and the issue comprises universal and language specific features. We found that ration of words of different length are dissimilar in individual texts, and the most common words are those composed of 5 phonemes and 2 syllables.

We evaluated word length in Tatar texts and attempted to fit a model based on Poisson distribution (in particular, a model based on one-displaced Poisson-uniform distribution was used), so description of empirical data was complemented with fitting theoretical values for word frequencies. *Besides, Shannon's entropy of word lengths was evaluated, and* a weak correlation between the average word length and entropy was found.

Keywords: word length, syllable, Poisson distribution, the Tatar language.

1 Introduction

Rigorously designed computational models can help making theoretical proposals by providing clues about their limitations or internal inconsistency, so they can contribute to improvement of theoretical approaches setting quite strict requirements for them. Length of linguistic items is one of significant formal features of languages which finds application in text processing, spell checking algorithms, language teaching, etc. Word length is studied by specialists in linguistics and text analysis as well as mathematicians and statisticians working on related issues. Basic approaches to word length are presented in papers by P. Grzybek [6], G. Altmann [1] and I. Popescu and colleagues [10].

Word length can be measured by the number of phonemes, morphemes or syllables in it, depending on research goals. A number of words of different length depends on the language type: for example, in synthetic languages, there is a higher morpheme-to-word ratio than in analytic languages [5]. The analytic structure of English (a great number of auxiliary verbs, prepositions as well as articles), determines existence of a great number of one-syllable words, so distribution of English word lengths can be

approximated by means of geometric distribution [2, 6]. The Tatar language uses agglutination to express syntactic relationships within a sentence, which, with lack of articles, significantly restricts the number of one-syllable words, limiting it to root words. So languages are characterized by dissimilar features in distribution of words of different length, which demands dissimilar ways of word length modeling.

Abundant linguistic data on word length provide great opportunities for selecting parameters and model fitting (Zipf´s law, Menzerath-Altmann law, approximating by means of different distributions, etc.). Behind a superficial simplicity of the concept of word length, a number of surprises may hide, "so here nothing helps but incessant testing, modeling, different viewing of data, modification of hypotheses, collecting of data from new languages, etc. Every "new" language can falsify a beloved theory or force us to modify it" [10].

There is a lack of special works devoted to word length in Turkic languages, although some data on Turkish is presented in overviews and papers covering multilingual data (for example, in [6, 7]. In dissertation by L. Rizvanova, certain aspects of word length in Tatar related to functional styles are considered [11]. On a special page of the Corpus of Written Tatar, a list of Tatar wordforms sorted by length is presented [13].

The aim of this paper is to empirically evaluate word length in Tatar and to model it using Poisson distribution. Tatar fiction texts are used as empirical source; the written texts were brought into a phonologically relevant form, to allow counting the number of phonemes and syllables per word.

2 Word length in Tatar texts

We examined a distribution of words of different length in 10 Tatar fiction texts (prose and poetry texts were used, 5 of both kind; brief information on selected texts is presented in Appendix). The written texts were brought into the standard form: 1 letter – 1 sound, and special rules were set to convert Tatar texts into a phonologically relevant form. When tokenizing, co-compounds (like ata-ana ('mother' + 'father') 'parents' [4] were regarded as individual words.

Figure 1 represents the number of words of different length in the text by A. Eniki; the word lengths are measured in phonemes. This text contains 2,169 words with length from 1 to 14 phonemes. Words with length 5 are the most frequent.

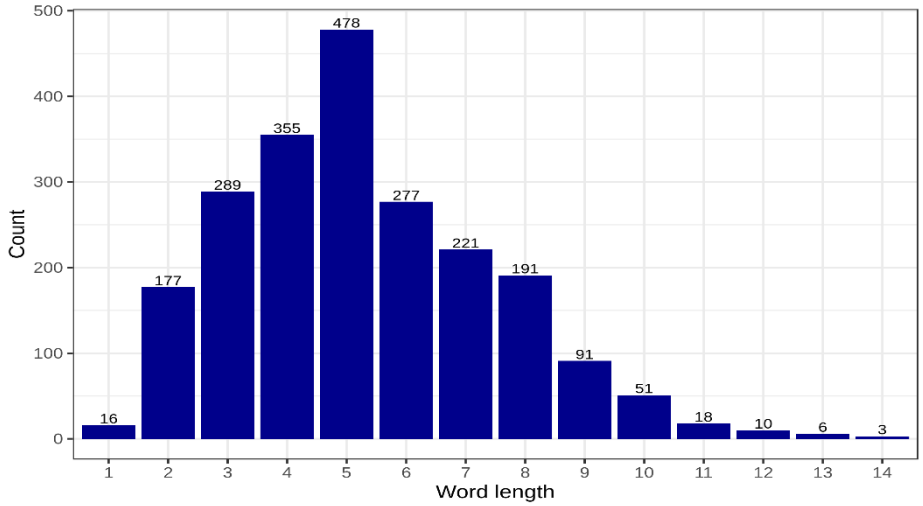

Fig. 1. Word frequencies by word length (measured in phonemes) in the text by A. Eniki

Words in the text by A. Eniki contain 4,962 syllables; the words are composed of one to six syllables, with the most frequent words being those with two syllables (see Figure 2).

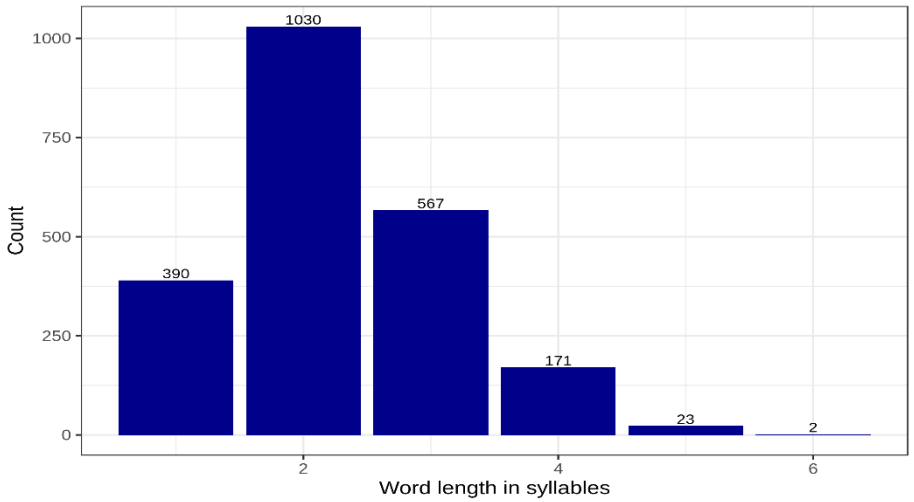

Fig. 2. Word frequencies by word length (measured in syllables) in the text by A. Eniki

The distribution of words depending on the number of phonemes and the number of syllables is represented in Figure 3. The most frequent are words consisting of 4 and 5 phonemes divided into 2 syllables.

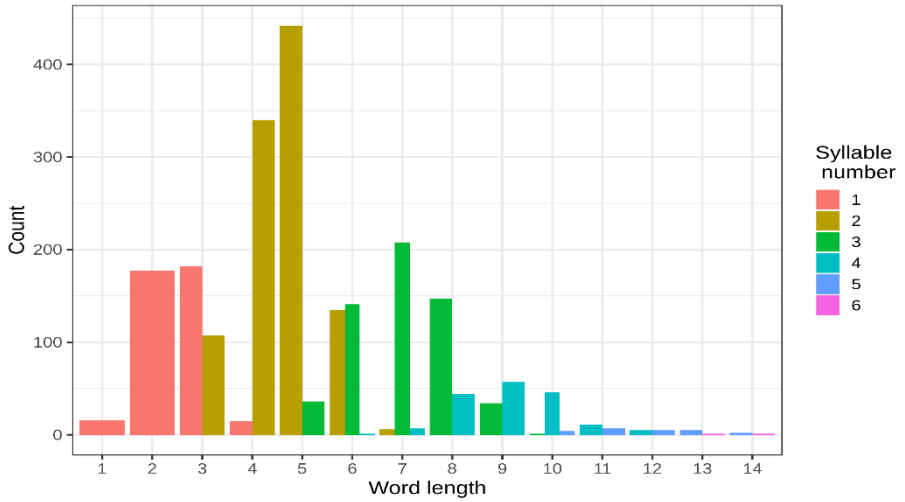

Fig. 3. Frequencies of words by their length in phonemes and in syllables in the text by A. Eniki

We computed the ratio of words consisting of a different number of phonemes in the selected texts, then calculated the entropy [12] of word length; the results are represented in Table 1.

Table 1. Relative frequencies of words of different length, mean word length and entropy of word length in Tatar texts

WL in phonemes	**Portion in texts**									
	Eniki	**Tukay, *Şüräle***	**Tukay, *Su anası***	**Tukay, *Käcä belän sarık***	**Amir-khan**	**Ibra-himov**	**Alish**	**Gilman**	**Suley-man**	**Zulfat**
1	0.007	0.006	0.019	0	0.005	0	0.004	0.005	0.003	0
2	0.081	0.082	0.103	0.071	0.064	0.038	0.074	0.08	0.073	0.031
3	0.132	0.195	0.179	0.1	0.128	0.112	0.126	0.132	0.124	0.117
4	0.163	0.121	0.098	0.25	0.128	0.119	0.144	0.145	0.144	0.196
5	0.219	0.231	0.27	0.321	0.201	0.238	0.217	0.206	0.22	0.307
6	0.127	0.13	0.129	0.107	0.13	0.117	0.136	0.143	0.166	0.153
7	0.101	0.082	0.084	0.038	0.133	0.13	0.117	0.112	0.068	0.131
8	0.087	0.086	0.069	0.033	0.093	0.119	0.083	0.07	0.045	0.061
9	0.042	0.027	0.019	0.019	0.058	0.047	0.039	0.056	0.09	0.049
10	0.023	0.018	0.017	0.038	0.031	0.054	0.033	0.025	0.051	0.037
11	0.008	0.015	0.014	0.017	0.024	0.025	0.016	0.011	0.017	0.012
12	0.005	0.003	0	0.005	0.004	0.002	0.007	0.006	0	0.006
13	0.003	0.001	0	0	0.002	0	0.001	0.004	0	0
14	0.001	0.001	0	0	0	0	0.001	0.002	0	0
15	0	0	0	0	0	0	0	0.001	0	0

16	0	0	0	0	0	0	0	0.002	0	0
17	0	0	0	0	0	0	0.001	0	0	0
Mean WL	5.291	5.103	4.909	5.024	5.65	5.897	5.478	5.443	5.53	5.374
Entropy	3.101	3.021	2.971	2.742	3.177	3.087	3.139	3.172	3.084	2.827

We found a weak correlation between the mean word length and the entropy: the correlation coefficient is 0.57. According to the sample of the texts, the average entropy of word length in prose is greater, and is less significant in poetry: it makes 2.923 for poetical texts and 3.135 for prose. This suggests that issues related to entropy in Tatar texts require further study.

3 Towards modeling

A great variety of distributions has been tested in word length studies, and Poisson distribution is often in focus of researchers who can modify it in different ways. Poisson distribution is a very simple, and in many cases sufficient means for modeling. Researchers believe that Poisson distribution, either in its usual form, or displaced to the right, or truncated above the zero point (positive Poisson distribution) should be used at the very beginning of any investigation [10]. In particular, the one-displaced Poisson-uniform distribution was used by V. Kromer to model German texts [8].

We follow this way and rely upon the approach by V. Kromer [7, 8] for modeling word length in Tatar. V. Kromer proposed a mathematical model of word length based on the Čebanov-Fucks distribution [3, 9] with equal distributions of the parameter. The Čebanov-Fucks distribution is a modification of the known Poisson distribution when the obligatory (first) syllable is not taken into account:

$$P_x = (\lambda - 1)^{x-1} / (x-1)! * e^{-(\lambda-1)}, x = 1, 2, 3, \ldots, \quad (1)$$

where Px is probability of textual word occurrence with length x, and λ is the distribution parameter [7, 8]. The latter could be estimated by the mean word length in the text (λ0), so this parameter is strictly determined by the text data.

We computed the probabilities of the textual word occurrence with the given length and obtained theoretical values for occurrences of words with the given length in each text. The results are given in Table 2 (fitted values with λ_0 parameter).

Then parameter λ_0 was replaced by lambdas belonging to the interval: [λ_0 - 0.1 * λ_0, λ_0 + 0.1 * λ_0]; length out was 150 for best approximation, and theoretical values of word occurrences for each case were computed. Then the discrepancy between experimental and theoretical data was evaluated by the Pearson Chi-square criterion χ^2 and the best fitted values were selected (λ_*). The results for 5 texts are represented in Table 2.

Table 2. Word occurrences with given lengths: observed and fitted values

WL in syllables	Eniki			Tukay, *Şüräle*			Tukay, *Su anası*			Tukay, *Käcä beläп sarık*			Suleyman		
	Observed	Fitted, λ_0	Fitted, λ_*	Observed	Fitted,λ_0	Fitted, λ_*	Observed	Fitted, λ_0	Fitted, λ_*	Observed	Fitted, λ_0	Fitted, λ_*	Observed	Fitted, λ_0	Fitted, λ_*
1	390	611	657	249	317	308	114	148	143	93	194	188	63	62	90
2	1030	778	789	422	339	339	198	154	154	395	212	212	179	108	124
3	567	495	474	196	182	186	84	80	82	36	116	119	58	94	85
4	171	210	190	54	65	68	23	28	29	55	42	44	55	55	39
5	23	67	57	4	17	19	0	7	8	0	11	12	0	24	13
6	2	17	14	0	4	4	0	1	2	0	3	3	0	8	4
λ		2.273	2.201		2.072	2.099		2.038	2.072		2.092	2.122		2.296	2.3712
χ^2		401.6	374.7		80.4	51,0		29.5	29.8		283.7	282.9		92.5	65.2
Total		2183			925			419			579			355	

Table 2 data evidence that occurrences of one-syllable words are overrepresented and occurrences of two-syllable words are underrepresented when fitting for all the texts. So using other modifications of Poisson distribution as well as using other distributions is needed in further research to achieve better fitting results.

4 Conclusion

We empirically evaluated word length in Tatar texts and attempted to fit a model based on Poisson distribution. Word lengths were measured in terms of phonemes and in terms of syllables.

Texts are not homogeneous because of internal rules of self-organization, so portions of words of different length are dissimilar in individual texts, and the most common words are those composed of 5 phonemes and 2 syllables.

Shannon's entropy of word lengths was evaluated, and a weak correlation between the average word length and the entropy was found with correlation coefficient equaling 0.57. According to the examined sample of texts, the average entropy of word lengths in prose is greater, and less significant in poetry, which may be the case due to the requirements of the poetic meter.

For modeling lengths of Tatar words, the one-displaced Poisson-uniform distribution was used. Although the results are generally consistent with those for other languages described in literature, nevertheless there is a significant discrepancy between the observed and fitted values, so using other modifications of Poisson distribution as well as using other distributions is needed in further research.

The results of the study of word length in Tatar can help in development of applications for style and register detection, authorship analysis and language teaching as well as they can be used in theoretical studies of language structure and complexity.

5 Acknowledgments

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

References

1. Altmann G.: Aspects of word length. Issues in Quantitative Linguistics, 3, 23–38. RAM-Verlag, Lüdenscheid (2013).
2. Elderton, W.P.: A Few Statistics on the Length of English Words. Journal of the Royal Statistical Society, series A (general), 112, 436–445. Wiley, New Jersey (1949).
3. Fucks, W.: Matematicheskaja teorija slovoobrazovanija [Mathematical theory of word formation]. In: Teorija peredachi soobshhenij (Trudy 3 mezhdunarodnoj konferencii) [Theory of messaging (3rd conference proceedings)], 221–247. Foreign Languages, Moscow (1957).
4. Galieva A., Suleymanov D.: Tatar Co-compounds as a Special Type of Classifiers. In: Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, 678–684. Ivane Javakhishvili Tbilisi State University, Tbilisi (2016).
5. Greenberg, J.: A Quantitative Approach to the Morphological Typology of Language. International Journal of American Linguistics, 26(3), 178–194. The University of Chicago, Illinois (1960).
6. Grzybek, P.: History and methodology of word length studies. In: Contributions to the Science of Text and Language. Word Length Studies and Related Issues, 15–90. Springer, Heidelberg (2005).
7. Kromer, V.: About Word Length Distribution. In: Contributions to the Science of Text and Language. Word Length Studies and Related Issues, 199–210. Springer, Heidelberg (2005).
8. Kromer, V.: Word length model based on the one-displaced Poisson-uniform distribution. Glottometrics, 1, 87–96. RAM-Verlag, Lüdenscheid (2001).
9. Piotrovskij, R.G., Bektaev, K.B., Piotrovskaja, A.A.: Matematicheskaja lingvistika [Mathematical linguistics]. High School, Moscow (1977).
10. Popescu, I.I., Naumann, S., Kelih, E., Rovenchak, A., et. al: Word length: aspects and languages. Issues in Quantitative Linguistics, 3, 224–281. RAM-Verlag, Lüdenscheid (2013).
11. Rizvanova, L.M.: Kvantitativnaja harakteristika tatarskogo slova: na materiale otdel'nyh funkcional'nyh stilej [Quantitative features of the Tatar word: on material of functional styles]. Candidate thesis in filology. Kazan University, Kazan (1996).
12. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal 27(3), 379–423. Wiley, New Jersey (1948).
13. Statistics. Corpus of Written Tatar, www.corpus.tatar/index_en.php?of=stat_en.htm, last accessed 2020/11/13.

Appendix

Basic information on the texts processed

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