УДК 809.437.5

COMPARATIVE STUDY OF QUANTITATIVE UZBEK FOLKLORE TEXTS

D.B. Urinbaeva

Samarkand Affiliate of Uzbekistan Academy of Sciences;
dilbarxon@inbox.ru

Represented by a Member of the Editorial Board Professor V.I. Konovalov

Key words and phrases: the average frequency of words; fill factor; epos; folklore; frequency; quantitative; sampling; rate synthetism; statistics.

Abstract: The article deals with the lexical and statistical structure of Uzbek folklore texts on the basis of quantitative comparative research: the average frequency of word forms, text coverability with different parts of the frequency lexicon, and the ratio of rare (random) lexical units.

1. Statement of the problem

Quantitative linguistics, together with other linguistic disciplines is involved in the task of constructing a theory of language. Cognition is not the main objective of quantitative linguistics; its primary goal is in finding necessary tools and techniques. Linguistics requires some ordinal and metric, i.e. quantitative concepts, and at the same time models and methods, which are naturally based on qualitative terms.

As part of the quantitative-typological study of the lexicon we are interested in the possibility of classifying the lexicon on the basis of grammatical features and the identification of the distribution of derived classes in the dictionary. In this paper we study and review lexical and grammatical classes of words - parts of speech. We have attempted to research the parts of speech of folklore texts by means of statistical analysis.

The starting material for our study is based on observations of five genres of folklore, such as epos (poem) [1], Uzbek folk tales [2], riddles [3], proverbs [4] and songs [5]. In this research we have compiled some frequency dictionaries for each of these genres, and then combined them into a single dictionary, in which the frequency of usage was indicated.

The first task was to get a general idea of the statistical characteristics of the lexicon of folklore texts. The total volume of texts is equal to 60 358 words. In particular, the total amount of words in the genre of epos reaches 14 029 units with the frequency of 96 011, which accounts for 49.2 % percent of the total. The total volume of fairy tales is equal to 14 837 with the frequency of 77 304 (24.7 %). The total amount of songs is 10 692 words with the frequency of 31 858 (27.6 %). The total volume of riddles is 8569 with the frequency of 27 334 (23.5 %). And finally, the total amount of proverbs is 12 231 words with the frequency of 45 132 (27.5 %).

For our research we established the length of the sample, which first was equal to 1000 words, then this sample was doubled to 2000 words, afterwards it reached 6000 words , and finally, the total amount of folklore texts

2. The coefficient of synthetism and the average frequency of word forms in Uzbek folklore texts

The resulting experimental frequencies of the individual word forms are random variables. This means that specific values of these quantities depend both on the set of those random factors and the situations in which this private sample of texts was created and realized as well as the quantitative experiment itself.

The problem of comparing the lexicon of the texts is discussed in the majority of papers devoted to the issues of the writer's lexicon and the history of literary language, because such works generally contain statements about the similarities or differences in the wordlist of authors or texts. Similar problems arise in other areas of theoretical and applied linguistics. Apparently, it was interesting to compare the vocabulary of different authors or styles. However, currently we are interested in the study of folklore texts vocabulary, largely due to the presence of unique folklore materials. This enables to set various tasks related to the application of statistical methods to the study of language. As a quantitative typological criterion we can use: firstly, the comparison of $\tilde{F}$ values by genres, and secondly, the rate of $\tilde{F}$ growth rates as a function of increasing sample size (in our case the initial sample was 1000 word forms). Tables 1–3 show the average frequency of word forms in various genres, obtained by processing approximately the equal length of samples.

When the amount of word usage is 1000 words in the genres of epos, proverbs and riddles the average frequency of word forms $\tilde{F} = 1.7$, in fairy tales $\tilde{F} = 1.8$, in the songs of $\tilde{F} = 1.5$.

When the volume reaches 2000 and 6000 word usages in folk songs genre the average repeatability of word forms is identical and equals $\tilde{F} = 1.6$ (2–3-tables). In riddles this index is higher and makes in both word usages (2000 and 6000) $\tilde{F} = 2.4$.

Table 1

Average repeatability of word forms and synthetism coefficient in samples of 1000 word usage

Genres	N	L_{wf}	$\tilde{F}$	Sint
Epos	1000	560	1.7	56
Fairy tales	1000	553	1.8	55
Proverbs	1000	582	1.7	58
Fork songs	1000	627	1.5	62
Riddles	1000	559	1.7	55

Table 2

Average repeatability of word forms and synthetism coefficient in samples of 2000 word usage

Genres	N	L_{wf}	$\tilde{F}$	Sint
Epos	2000	1036	1.9	51
Fairy Tales	2000	925	2.1	46
Proverbs	2000	1097	1.8	54
Fork Songs	2000	1208	1.6	60
Riddles	2000	822	2.4	41

Table 3

Average repeatability of word forms and synthetism coefficient in samples of 6000 word usage

Genres	N	L_{wf}	$\tilde{F}$	Sint
Epos	6000	2685	2.2	44
Fairy Tales	6000	2499	2.4	41
Proverbs	6000	2898	2	48
Fork Songs	6000	3725	1.6	62
Riddles	6000	2413	2.4	40

The average repeatability of word forms in the total amount of the folklore texts is equal to (see Table 4): $\tilde{F} = 2.9$ (folk songs), $\tilde{F} = 6.8$ (epos), that testifies to higher variety of words and grammatical forms of national songs in relation to folk epos. Dynamics of growth of this index can be observed in the following table.

Generalization of the value of synthetism factor in Uzbek folklore texts shows the variety of their values. When comparing the samples of 1000, 2000, 6000 words (see Table 5), and the total of all the material the highest synthetism coefficient is observed in folk songs (33), and the lowest is in epos (15).

It should be noted that different rates of increase in the average frequency of word forms, depending on the increase in the sample size change not only from language but also from style to style.

Table 4

Average repeatability of word forms and synthetism coefficient in folklore texts

Genres	N	L_{wf}	$\tilde{F}$	Sint
Epos	96011	14029	6.8	15
Fairy Tales	77304	14837	5.2	19
Proverbs	45135	12231	3.6	27
Fork Songs	31858	10692	2.9	33
Riddles	27334	8569	3.1	31

Table 5

Dynamics of growth of average repeatability of word forms and synthetism factor

Frequency of lexicon and word forms	N (word usage)							
	Zones						Total volume	
	1–1000		1–2000		1–6000			
Epos	1.7	56	1.9	51	2.2	44	6.8	15
Fairy tales	1.8	55	2.1	46	2.4	41	5.2	19
Proverbs	1.7	58	1.8	54	2	48	3.6	27
Fork songs	1.5	62	1.6	60	1.6	62	2.9	33
Riddles	1.7	55	2.4	41	2.4	40	3.1	31

3. The occupancy rate of the text with commonly used word forms

Uzbek grammatical system, as well as the general system of the Turkic languages, can generate a virtually infinite number of word forms. In this regard, in the study of lexical and statistical structure of the Uzbek folklore text it is especially important to use a comparative study of the average frequency of word forms, text coverability different parts of the frequency dictionary, and the ratio of rare and frequent lexical items.

In solving a number of theoretical and applied problems it is necessary to select the vocabulary, which has a sufficiently high probability to occur at random texts of the language, style, or sub-language. Based on this sample text, we should separate rare word forms and words from the lexical items that have medium to high usage, and then determine what percentage of them cover these medium-and high-used words. Based on the experience of predecessors, the borderline between rare and commonly used lexical units is determined; it can be drawn between units used once or twice.

We consider the evaluation of the weight of rare word forms in terms of quantitative typology, suggesting that they can serve as the characteristics that distinguish the texts of folklore. To do this, when comparing the infrequent word forms, we will rely on a sample of the same volume for the texts of folklore in Tables 6–9, where the comparison of rarely used word forms is presented in 1000, 2000, 6000 sample sizes in various genres of folklore.

At the 1000 sample the usage of rarely used word forms in the texts of a fairy tale reaches $\xi = 59$, while in folk songs it reaches $\xi = 49$. With the increase of sample size to 2000 word forms the share of rarely used word forms in fairy tales has increased and reached $\xi = 66$, and in the proverbs $\xi = 47$.

With increase of sample size to 6000 word forms the share of rarely used word forms in proverbs equals $\xi = 76$, in folk songs it accounts for $\xi = 62$. With the increase in the sample size the frequency of rarely used word forms has increased as well.

Table 6

Quantitative data on samples in 1000 word forms

Frequency the dictionaries and word forms	N	L_{wf}	F_1	$F_{1\%}$	F_2	$F_{2\%}$	$F_{1,2}$	$F_{1,2\%}$	ξ
Epos	1000	553	F_{385}	38	F_{73}	7.3	F_{458}	45.3	55
Fairy tales	1000	500	F_{346}	34.6	F_{67}	6.7	F_{413}	41.3	59
Proverbs	1000	582	F_{435}	43.5	F_{72}	7.2	F_{507}	50	50
Fork songs	1000	559	F_{459}	46	F_{51}	5.1	F_{510}	51	49
Riddles	1000	627	F_{371}	37	F_{83}	8.3	F_{454}	45.4	55

Table 7

Quantitative data on samples in 2000 word forms

Frequency the dictionaries and word forms	N	L_{wf}	F_1	$F_{1\%}$	F_2	$F_{2\%}$	$F_{1,2}$	$F_{1,2\%}$	ξ
Epos	2000	1090	F_{774}	38.7	F_{158}	7.9	F_{932}	46.6	54
Fairy tales	2000	856	F_{539}	26.9	F_{144}	7.2	F_{683}	34.1	66
Proverbs	2000	1208	F_{898}	44.9	F_{167}	8.3	F_{1065}	53.2	47
Fork songs	2000	1197	F_{828}	41.4	F_{202}	10	F_{1030}	51.5	49
Riddles	2000	1036	F_{697}	34.8	F_{170}	8.5	F_{867}	43.3	57

Table 8

Quantitative data on samples in 6000 word forms

Frequency the dictionaries and word forms	N	L_{wf}	F_1	$F_{1\%}$	F_2	$F_{2\%}$	$F_{1,2}$	$F_{1,2\%}$	ξ
Epos	6000	2499	F_{1607}	26,7	F_{396}	6,6	F_{2003}	33	67
Fairy tales	6000	2615	F_{1744}	29	F_{358}	5,9	F_{2102}	35	65
Proverbs	6000	2725	F_{1875}	31,2	F_{419}	6,9	F_{1456}	25	76
Fork songs	6000	2693	F_{1923}	32	F_{408}	6,8	F_{2331}	38	62
Riddles	6000	2685	F_{1775}	29	F_{419}	6,9	F_{2194}	36	64

Table 9

Quantitative data

Frequency the dictionaries and word forms	N	L_{wf}	F_1	$F_{1\%}$	F_2	$F_{2\%}$	$F_{1,2}$	$F_{1,2\%}$	ξ
Epos	96011	14029	F_{7404}	52,7	F_{4342}	30,9	F_{11746}	83,6	88
Fairy tales	77304	14837	F_{8265}	55,7	F_{4580}	30,8	F_{12845}	86,5	83
Proverbs	45132	12231	F_{7029}	57,4	F_{4054}	33,1	F_{11083}	90,5	75
Fork songs	31858	10692	F_{6809}	63,6	F_{2538}	23,7	F_{9347}	87,3	71
Riddles	27334	8569	F_{5275}	61,5	F_{2636}	39,7	F_{7911}	92,2	72

Table 10

Dynamics of growth of rarely used word forms

Frequency the dictionaries and word forms	N (word using)			
	Zones			Total volume
	1–1000	1–2000	1–6000	
Epos	55	54	67	88
Fairy tales	59	66	65	83
Proverbs	50	47	76	75
Fork songs	49	49	62	71
Riddles	55	57	64	72

In the total amount of samples the highest frequency is shown in the texts of epos $\xi = 88$, the lowest rate is shown by folk songs $\xi = 71$.

From the above it follows that a statistical experiment conducted under identical conditions on different folklore genres gives results that vary on reliability and quality. We can observe the growth dynamics of these indicators, however, in fairy tales with an increase in sample size the share of infrequent word forms has remained almost the same.

In the total sample we give a brief quantitative and typological commentary between genres of folklore.

The behavior of the filling factor ξ between the genres of folklore in total as follows:

Epos-fairy tales: the value of ξ in epos is 5 % more than in a fairy tale ($\xi_{\text{fairytales}} < \xi_{\text{epos}}$).

Epos-proverb: the value of ξ in epos is 13 % more than ($\xi_{\text{proverbs}} < \xi_{\text{epos}}$).

Epos-folk songs: the value ξ in epos is 17 % bigger than folk songs ($\xi_{\text{folksongs}} < \xi_{\text{epos}}$).

Epos-riddle: the value of ξ in epos is 16 % bigger than in riddles ($\xi_{\text{riddles}} < \xi_{\text{epos}}$).

The value is lower in folk songs by 17 % in comparison with other genres; that testifies to the prevalence of rarely used word forms.

Comparing of value of the filling factor in the texts containing rarely used word forms in the examined genres of the Uzbek folklore text, we receive the following imbalance:

$$\xi_{\text{folksongs}} < \xi_{\text{riddles}} < \xi_{\text{proverbs}} < \xi_{\text{fairytales}} < \xi_{\text{epos}}.$$

This imbalance is a parameter of distinction in a variety of compared genres of folklore. Thus, the lower the value of filling factor ξ, the more varied is the examined genre. In our case such a genre is folk songs.

4. Conclusions

We have investigated the frequency dictionaries; it is the simplest and at the same time useful way to describe the vocabulary of statistics. The total consideration of diverse structure and objectives of the frequency dictionaries shows that in addition to a number of specific issues related to compiling dictionaries of a specific type, it is necessary to solve a more general problem, namely, to provide a technique for compiling a frequency dictionary so as to receive the information with the required accuracy of a given percentage of words of text.

The quantitative study of Uzbek folklore texts and the comparison of the data make it possible to reveal quantitative typological differences between the studied texts. These differences are consistently and unequivocally found in such quantities as the average frequency of word forms, the growth of statistical coverability of the text, occupancy of the text with the most widely used and rarely used word forms.

In ours linguistic statistical experiment the value of the synthetism coefficient in the examined genres of Uzbek folklore texts can be expressesd in the following inequality

$$\text{Sint}_{\text{epos}} < \text{Sint}_{\text{fairytales}} < \text{Sint}_{\text{proverbs}} < \text{Sint}_{\text{riddles}} < \text{Sint}_{\text{folksongs}}.$$

As this inequality shows folk songs have the highest rate, while epos has the lowest. It testifies that the folk songs have dominant statistical weight from the point of view of lexicon variety and their grammatical forms in relation to other genres of folklore.

We also used filling factor of the text with rarely used lexical units to compare various genres of folklore. The comparison of these factors by genres of folklore, can be shown the following inequality

$$\xi_{\text{folksongs}} < \xi_{\text{riddles}} < \xi_{\text{proverbs}} < \xi_{\text{fairytales}} < \xi_{\text{epos}}.$$

This inequality also testifies to a variety of the texts of folk songs.

The combination of informational and statistical experiment enables to distinguish typological distinctions among genres of folklore. In folk songs the filling factor is the highest that obviously testifies to their greater analytics in comparison with other genres. The latter observation is consistent with the specific historical destinies of genres such as folk songs that are rich in children's folklore (take for instance those we've investigated). Children's folklore is one of the most lively and rich phenomena of Uzbek culture. It combines both ancient and contemporary works. And they both are continuously updated and redone. Both adults' folklore and children's folklore reflect the history. Funny poems, merry songs, amusing teasers, twisters are passed by children to each other. This is a national children's art expression. In the dictionary of folk songs, a lot of obscure words have been invented by children. Due to this the lexis of folk songs is more varied compared to other genres.

References

1. Алпомиш. Фозил Йўлдош ўғли. – Тошкент : «Шарқ» нашриёти-матбаа концерни бош таҳририяти, 1998.
2. O'zbek xalq ertaklari. 1 tom. – Toshkent : "O'qituvchi" nashriyot-matbaa ijodiy uyi, 2007.
3. Топишмоқлар. – Тошкент : Ғ. Ғулом номидаги Адабиёт ва санъат нашриёти, 1981.
4. O'zbek xalq maqollari. – Toshkent : «Sharq» nashriyot-matbaa aksiyadorlik kompaniyasi bosh tahririyati, 2005.
5. Бойчечак. – Тошкент : Ғ. Ғулом номидаги Адабиёт ва санъат нашриёти, 1984.
6. Айимбетов, М.К. Проблемы и методы квантитативно-типологического измерения близости тюркских языков : автореф. дис. … д-ра филол. наук : 10.02.06 / М.К. Айимбетов. – Ташкент, 1997. – 43 с.
7. Садыков, Т. Проблемы моделирования трюкской морфологии / Т. Садыков. – Фрунзе : Илим, 1987. – 120 с.
8. Тулдава, Ю. Проблемы и методы квантитативно-системного исследования лексики / Ю. Тулдава. – Таллин : Валгус, 1987. – 204 с.
9. Мухаммедов, С.А. Инженерная лингвистика и опыт системно-статистического исследования узбекских текстов / С.А. Мухаммедов, Р.Г. Пиотровский. – Ташкент : Фан, 1986. – 163 с.

Сопоставительные квантитативные исследования узбекских фольклорных текстов

Д.Б. Уринбаева

Самаркандское отделение Академии наук Республики Узбекистан;
dilbarxon@inbox.ru

Ключевые слова и фразы: выборка; дастан; квантитатив; коэффициент заполнения; коэффициент синтетизма; средняя повторяемость слова; статистика; фольклор; частота.

Аннотация: Раскрыта лексико-статистическая структура узбекского фольклорного текста на основе сопоставительного квантитативного исследования: средней повторяемости словоформ, покрываемости текста различными участками частотного словаря, а также соотношения редких (случайных) лексических единиц.

Vergleichende quantitative Untersuchung der usbekischen Folkloretexte

Zusammenfassung: Es ist die lexikalisch-statistische Struktur des usbekischen Folkloretextes auf Grund der vergleichenden Quantitativuntersuchung: der mittleren Reproduzierbarkeit der Wortformen, der Deckbarkeit des Textes von den verschiedenen Bereichen des Häufigkeitswörterbuches und auch der Korrelation der seltenen Lexikaleinheiten eröffnet.

Etude comparative quantitative des textes folkloriques ouzbeks

Résumé: Est montrée la structure lexico-statistique du texte folklorique ouzbek à la base de la recherche comparative quantitative de la répétition moyenne des formes des mots, de la présénce de différents secteurs du vocabulaire de fréquence ainsi que de la relation des unités lexicales rares (occasionnelles).

Автор: ***Уринбаева Дилбар Базаровна*** – кандидат филологическх наук, заместитель директора Самаркандского отделения Академии наук Республики Узбекистан.

Рецензент: ***Юлдашев Б.*** – доктор филологических наук, профессор Самаркандского государственного университета.
