International Journal Of Literature And Languages

VOLUME Vol.05 Issue05 2025
PAGE NO. 85-90
DOI 10.37547/ijll/Volume05Issue05-24

Determination of High-Frequency Wordforms In Fitrat's Works Using Corpus

Alimbekova Mavjuda Xalimjon qizi
Phd Student at Tashkent State University of Uzbek Language and Literature, Uzbekistan

Received: 29 March 2025; **Accepted:** 25 April 2025; **Published:** 27 May 2025

Abstract: This article describes the semi-automatic determination of the frequency of wordforms used in the works of Abdurauf Fitrat and their statistics. High frequency words were analyzed within word groups. From the hundred highest frequency words, words with lexical meaning were studied, which help to identify the leading themes in the author's work.

Keywords: Abdurauf Fitrat corpus, wordform frequency, independent word, high frequency word, wordform statistics.

Introduction: The development of computer technology makes it possible to analyze the language of an author's works using a corpus. Research dedicated to language studies has begun to be conducted based on corpus-based approaches. Studying the linguistic features and vocabulary of Fitrat's works and compiling dictionaries are among the important tasks of linguistics. The creation of the Abdurauf Fitrat corpus not only helps preserve and compile the author's works in one place but also facilitates the study of the language used in his writings. Creating a dictionary of Fitrat's works, analyzing his stylistics, studying the usage period and frequency of words—all highlight the significance of developing a corpus of Fitrat's texts. This allows researchers to explore the author's style and word usage skills, analyze expressions and proverbs used by the author, study artistic devices and stylistic features in lyrical works, and better understand historical or obscure words and the overall idea of the text.

In this article, high-frequency words used in the author's works are statistically analyzed using a semi-automatic method. "Semi-automatic" refers to a mechanism that cannot function completely independently without human involvement and is not fully automated. In this study, word form frequencies were identified using a corpus, and their part-of-speech analysis was done manually. The statistical analysis was carried out in a "semi-automatic" manner, i.e., with the involvement of both corpus tools and human input. This approach contributes to the accuracy of the results obtained.

Literature Review

In global linguistics, there are numerous studies dedicated to analyzing the language of authors' works using corpora. The frequency-based grammatical-semantic dictionary of A.P. Chekhov's literary works created by O.V. Kukushkina, A.A. Polikarpov, and E.V. Surovseva [1] is a vivid example of such research. Several dictionaries, including a frequency dictionary and an idiomatic word database by A.Ya. Shaykevich [3], have been created based on Dostoevsky's author corpus [2].

In Uzbek linguistics, Sh. Hamroyeva developed the linguistic foundations of the Uzbek language author corpus and laid the foundation for the Abdulla Qahhor author corpus [4]. N. G'ulomova's study titled "The Author Corpus of Alisher Navoi and Its Semantic Tag Database (based on the 'Badoye' ul-vasat' collection)" [5] and the Alisher Navoi author corpus co-authored by M.A. Abjalova, N.S. G'ulomova, and Sh.M. Sa'dullayeva [6] serve as models for the works being created in this field.

Various specialists have conducted numerous studies on the literary heritage of Abdurauf Fitrat. The lexicon of Fitrat's works was first studied statistically and thematically by Y. Saidov in his dissertation titled "The

Lexicon of Fitrat's Literary Works" [7]. This research examined the lexical features of Fitrat's literary language and statistically analyzed native and borrowed words, as well as ancient Turkic elements used in the author's writings. The dissertation notes that Fitrat's dramas and poems were first critically analyzed by H. Olimjon. In the articles and research works by M. Qurbonova [8], including "Fitrat on Language Development," "Fitrat as a Linguist," and "Fitrat's Linguistic Legacy," and in B. To'ychiboyev's [9] articles such as "Fitrat and the Contemporary Uzbek Literary Language," the linguistic aspects and scholarly articles of Fitrat have been explored.

METHODOLOGY

In this article, the statistics and frequency of all word forms used in Abdurauf Fitrat's works were identified and semi-automatically analyzed using a corpus. The method of statistical analysis was employed. This method determines the repetition, frequency of use, and distribution scope of linguistic units.

RESULTS

A total of 75 works by Abdurauf Fitrat, covering 14 genres, were uploaded into the SketchEngine software. A small test corpus of Abdurauf Fitrat's works was created for analytical purposes. Using the corpus, 422,093 tokens and 58,643 word forms were identified (see Figure 1).

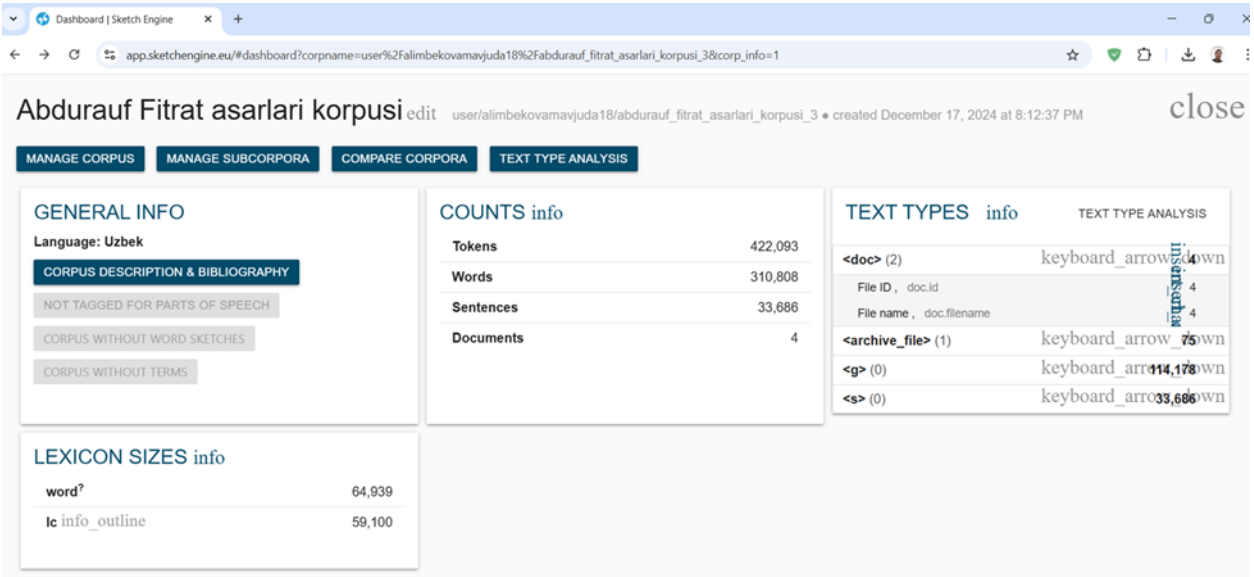

Figure 1. Test corpus created in Sketch Engine

According to word form frequency, the word bir ("one") is the most frequently used word, appearing 5,384 times. The conjunction va ("and") appears 4,312 times, and the pronoun bu ("this") occurs 4,192 times. A total of 11 words are used more than one thousand times. It was also found that 34,537word forms were used only once (see Figure 2).

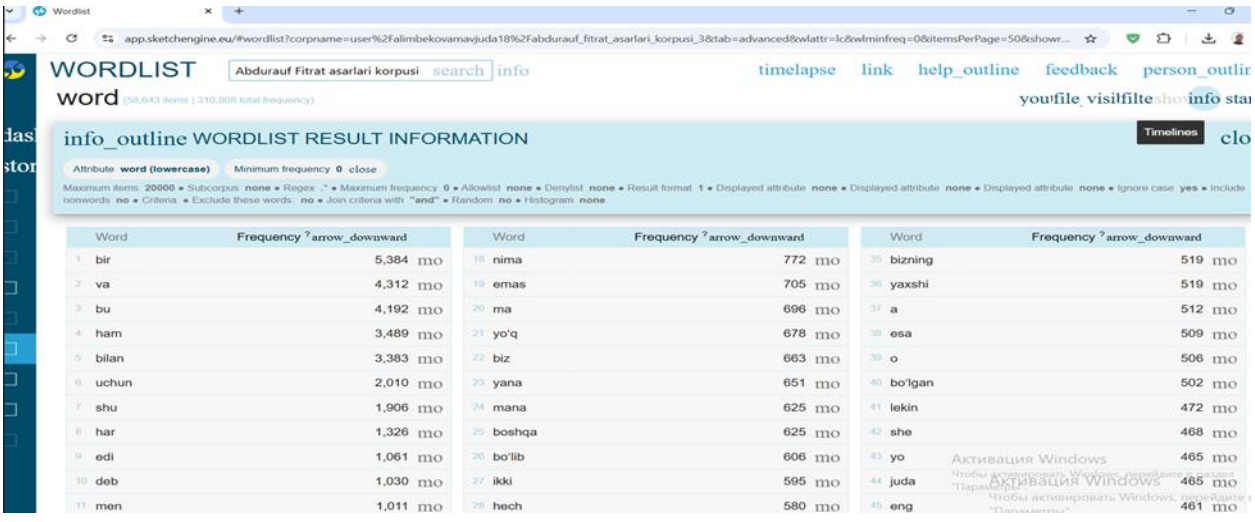

Figure 2. Word Form Frequency

When word forms were identified using this software tool, they were recognized based on their formal (morphological) structure. However, some shortcomings were observed—for instance, with words containing the letters ‘o‘, ‘g‘‘, and punctuation marks, as well as in the recognition of compound verbs. In the corpus texts, there are cases where the author has analyzed words by dividing them into syllables. In works dedicated to linguistics, suffixes are presented and analyzed separately. The SketchEngine software mistakenly treated such non-word syllabic forms and suffixes as independent words (see Figure 2). These incorrectly identified word forms were manually corrected. Words written in Arabic script and meaningless forms were deleted. A sample of 100 word forms and their frequencies can be seen in Table 1.

Table 1. ***List of the most frequent words***

Word Forms	Frequency	Word Forms	Frequency
bir	5384	shuning	383
va	4312	shul	376
bu	4192	ular	371
ham	3489	islom	365
bilan	3383	biroq	355
uchun	2010	degan	355
shu	1906	kishi	352
har	1326	qanday	352
edi	1061	sen	348
deb	1030	misra	347
men	1011	birinchi	344
bor	919	uch	339
qilib	878	odam	336
uning	842	butun	333
o‘z	840	bo‘ladir	332
u	826	Yuz	321
kabi	801	agar	317
nima	772	boy	315
emas	705	ish	313
yo‘q	678	narsa	312
biz	663	orasida	311
yana	651	lozim	302

boshqa	625	albatta	302
mana	625	qabul	293
bo‘lib	606	mumkin	291
ikki	595	oz	287
hech	580	katta	282
bo‘lsa	573	bizga	281
keyin	563	bo‘ldi	281
shunday	561	o‘n	273
bo‘lg‘an	547	ko‘ra	271
olib	545	mening	271
ekan	530	kelib	270
yaxshi	519	o‘zi	268
bizning	519	to‘g‘ri	264
esa	509	qiladi	263
bo‘lgan	502	biri	261
lekin	472	turli	258
yo	465	menga	256
juda	465	o‘zining	256
eng	461	tomonidan	256
siz	451	edilar	254
ularning	445	ul	250
kerak	434	qarab	247
kim	403	necha	244
uni	399	unga	243
ko‘b	398	meni	242
kun	397	so‘z	241
so‘ng	387	nega	240
bo‘ladi	385	buning	238

From the 100 most frequent word forms identified, we excluded auxiliary word classes, function words, pronouns, verbs, and numerals. We then extracted words that carry lexical meaning and reflect the themes of the author's creative works. These include: Islam (365), person (352), verse (347), human (336), rich (315), work (313), and word (241). The frequent use of these lexical units in the author's works indicates the dominant themes in Fitrat's writings.

Based on the words extracted from the word forms used in Fitrat's works (Table 1), we examined which content words were most frequently used:

1. Noun Word Class: kun (day) (397), islom (Islam) (365), kishi (person) (352), misra (verse) (347), odam (human) (336), yuz (face) (321), ish (work) (313), narsa (thing) (312), qabul (acceptance) (293);

2. Adjective Word Class: boshqa (other) (625), yaxshi (good) (519), katta (big) (282), butun (whole) (333), to'g'ri (correct) (264), turli (various) (258);

3. Numeral Word Class: bir (one) (5384), ikki (two) (595), birinchi (first) (344), uch (three) (339), o'n (ten) (273);

4. Pronoun Word Class: bu (this) (4192), shu (this) (1906), men (I) (1011), uning (his/her) (842), o'z (own) (840), u (he/she) (826), nima (what) (772), biz (we) (663), mana (here) (625), shunday (so) (561), bizning (our) (519), siz (you) (451), ularning (their) (445), kim (who) (403), uni (him/her) (399), shuning (that) (383), shul (that) (376), ular (they) (371), qanday (how) (352), sen (you) (348), mening (my) (271), o'zi (himself/herself) (268), menga (to me) (256), o'zining (his/her own) (256), ul (they) (250), necha (how many) (244), unga (to him/her) (243), meni (me) (242), so'z (word) (241), nega (why) (240), buning (this) (238);

5. Verb Word Class: edi (was) (1061), deb (said) (1030), qilib (doing) (878), emas (not) (705), bo'lib (being) (606), bo'lsa (if) (573), bo'lg'an (having been) (547), olib (taking) (545), ekan (was) (530), bo'lgan (happened) (502), bo'ladi (will be) (385), degan (said) (355), bo'ladir (will be) (332), bo'ldi (was) (281), kelib (coming) (270), qiladi (does) (263), edilar (they were) (254), qarab (looking) (247);

6. Adverb Word Class: keyin (then) (563), ko'b (many) (398), so'ng (after) (387), oz (few) (287).

Table: Out of the 100 word forms listed, 9 belong to the noun word class, 6 to the adjective word class, 5 to the numeral word class, 31 to the pronoun word class, 18 to the verb word class, and 4 to the adverb word class. It has been found that the most frequently used word class in the texts of Fitrat's works is pronouns. Following that, verbs and nouns are also used actively.

Among the auxiliary words, the most frequent word is va (and), which was used 4312 times, followed by ham (also) with 3489 occurrences. The word bilán (with) is used 3383 times, and uchun (for) appears 2010 times. The frequency of use of auxiliary words in the top 100 most frequent word forms is as follows:

1. Conjunctions: va (and) (4312), agar (if) (317), lekin (but) (472), bilán (with) (3383), yo (or) (465), biroq (however) (355);

2. Auxiliaries: uchun (for) (2010), deb (said) (1030), ko'ra (according to) (271), tomonidan (by) (256);

3. Predicatives: ham (also) (3489), har (every) (1326), kabi (like) (801), hech (no) (580), juda (very) (465), eng (most) (461).

The modal words belonging to the intermediary word class, such as bor (there is) (919), lozim (necessary) (302), albatta (certainly) (302), mumkin (possible) (291), and kerak (needed) (434), are also frequently used. However, exclamations and onomatopoeic words are not present in the top 100 most frequent words.

CONCLUSION

The statistics of word forms in Abdurauf Fitrat's works corpus, along with their frequency, were determined. The top 100 most frequent word forms were extracted and analyzed according to word classes. The most frequently used lexical words conveying significant meanings in the author's works were identified and studied. Studying such words helps in identifying the dominant themes in the author's creative works.

REFERENCES

Кукушкина О.В., Суровцева Е.В., Лапонина Л.В. Частотный грамматико-семантический словарь языка художественных произведений А.П.Чехова с электронным приложением. – М.: МАКС Пресс, 2012. – 571 с.

В.И.Заботкиной "Методи когнитивного анализа семантика слова компютерно-корпусной подход", Москва: Язика словянской култури, п. 348, 2015.

Словар язика Достоевского. Идеоглоссарий. // Российская академия наук институте руского язика им В.В.Виноградова, 2008.

Sh.M.Hamroyeva "O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari", Buxoro, – 259 bet, 2018.

N.S.G'ulomova "Alisher Navoiy mualliflik korpusi va uning semantik teglari bazasini yaratish ("Badoye' ul-vasat" devoni asosida)", Toshkent, – 190 bet, 2022.

http://navoiykorpusi.uz/

Саидов Ё. Фитрат бадиий асарлари лексикаси: Филология фанлари бўйича фалсафа доктори (PhD) дисс. –Samarqand, 2004.

Қурбонова М. Фитрат тил тараққиёти

ҳақида//Туркистон. – 1996. – 6 ноябр.

Қурбонова М. Фитрат – тилшунос. – Тошкент, 1996. – 29 б.

Қурбонова М. Фитратнинг тилшунослик мероси: Филол.фанлари номзоди… дис. Автореф. – Т., 1993.

Тўйчибойев Б. Фитрат ва ҳозирги ўзбек адабий тили// Фитрат анжумани материаллари. – Бухоро, 1992. – Б. 54-56.

https://SketchEngine