

World Conference on Social Sciences, Law and Public Policy
Hosted Online from Toronto, Canada
Date: 26th October, 2026
Website: https://econferencia.com

FREQUENCY OF USE OF THE CONCEPT "FRIENDSHIP" IN UZBEK AND RUSSIAN LANGUAGE CORPORA

Davanov Shamsiddin Faxriddin ugli
Independent Researcher at the Denov Institute of Entrepreneurship and Pedagogy

Abstract

This article examines the comparative frequency of the Uzbek lexeme "do'stlik" and the Russian lexeme "дружба" in language corpora. The study draws on corpus linguistics, quantitative analysis, and contextual-semantic methods. It outlines procedures for identifying raw and normalized frequencies, accounting for grammatical forms, examining genre and chronological distribution, and analyzing usage contexts. The Uzbek National Corpus and the Russian National Corpus are considered as electronic resources for comparative research. The article explains why raw counts from corpora of different sizes cannot be compared directly and why frequency values should be normalized. The proposed methodology provides a framework for investigating the lexical distribution of the concept of friendship in Uzbek and Russian.

Keywords: Corpus linguistics, concept of friendship, usage frequency, normalized frequency, national corpus, comparative analysis, contextual semantics.

Introduction

Corpus linguistics is a significant branch of modern linguistics that enables the collection, systematization, and quantitative analysis of natural language materials in electronic form. Corpora allow for the determination of how

frequently a specific lexeme is used, which words it collocates with, the grammatical forms in which it appears, and the contexts in which it is activated. Consequently, corpus analysis serves as an essential empirical tool for studying the behavior of lexical units in actual speech [1]. The concept of "friendship" (do'stlik) encompasses relationships between individuals characterized by trust, loyalty, sincerity, and mutual support. In the Uzbek language, the primary term for this concept is "do'stlik", while in Russian, it corresponds to the lexeme "дружба". Comparing the frequency of use of these units provides insight into how the concept of friendship is expressed within the lexical systems of the two languages. However, frequency metrics do not fully capture the significance of a concept within a culture. Results are also influenced by the composition of the corpus, the distribution of genres, the time period in which the texts were created, and the specific search methodology employed.

Analysis and results

The National Corpus of the Uzbek Language offers search capabilities based on word forms, lemmas, and various text categories. The Russian National Corpus also offers search tools for lexemes, word forms, and various sub corpora [2; 3]. These resources provide a methodological framework for the comparative study of the units "do'stlik" and "дружба" based on corpus data. The objective of this article is to develop principles for comparing the usage frequency of the lexemes "do'stlik" and "дружба" within Uzbek and Russian language corpora, to identify factors influencing this frequency, and to characterize their functional properties through contextual analysis.

In corpus research, a distinction is made between absolute frequency and relative frequency. Absolute frequency refers to the number of times a specific lexeme appears in a corpus, while relative frequency calculates the number of

occurrences in relation to the total size of the corpus. When comparing corpora of different sizes, relative frequency is a more suitable metric. Relative frequency per million tokens can be determined using the following formula: in this formula, (f) represents the number of occurrences of the target lexeme in the corpus, and (N) represents the total number of tokens in the corpus. The result indicates the number of occurrences per one million tokens.

For instance, the fact that a lexeme appears 500 times in one corpus and 800 times in another does not automatically mean it is used more frequently in the second language or context. If the second corpus is significantly larger than the first, the relative frequency might actually show the opposite result. Therefore, comparative analysis should include standardized frequency alongside absolute figures. In any such study, the unit of search must be clearly defined in advance. If the objective is to compare the specific lexemes “do‘stlik” and “дружба”, the frequency of these words should first be determined based on their lemmas. In the next stage, their case or number forms, as well as derived units, may be considered separately. For example, forms such as “do‘stlikni”, “do‘stlikdan”, and “do‘stlikka” in Uzbek, or “дружбы”, “дружбе”, and “дружбу” in Russian, can either be grouped under the relevant lemma or counted separately. Whichever approach is chosen; it must be clearly specified in the research methodology.

Furthermore, related or derivative units - such as “do‘st”, “друг”, “do‘stona”, and “дружеский” - cannot simply be automatically added to the primary frequency count. While such units are useful for exploring the broader lexical field of a concept, they must be calculated separately from the frequency of the core lexeme. Regarding the lexeme “do‘stlik” (friendship) in the Uzbek language: to determine its frequency of use, one can perform a search by word form or lemma within the search system of the National Corpus of the Uzbek

Language. The corpus's search interface allows for filtering by text style and time period [2]. This feature is essential for studying the distribution of a lexeme across different layers of text.

Depending on the context, the lexeme "дружба" (friendship) can denote interpersonal relationships, collective solidarity, ties between nations, or political-diplomatic cooperation. Therefore, relying solely on a quantitative count of search results is insufficient. Contexts retrieved from the corpus should be categorized into semantic groups to examine which specific meanings are activated within particular types of text. The size and composition of various sub-corpora within the National Corpus of the Russian Language may differ from one another. For instance, the main corpus, media texts, and social media materials represent different communicative situations [4]. Therefore, not only the overall frequency but also the relative frequency within a selected sub corpus must be taken into account.

When comparing overall frequency, both absolute and relative figures for the two lexemes are determined. In this process, the size and composition of each corpus must be taken into account. If the corpora differ significantly in terms of size, time period, or the proportion of genres, the results cannot be directly interpreted as inherent linguistic characteristics. In such cases, additional analysis should be conducted using smaller, yet comparable, sub-corpora drawn from similar genres and time frames. Genre-based analysis helps identify the communicative contexts in which specific lexemes appear most frequently. For instance, meanings related to personal relationships may be more prominent in literary texts, whereas meanings concerning collective or international relations might be more prevalent in journalistic or publicity texts. However, such differences are not predetermined conclusions, but rather scientific hypotheses to be tested against corpus data.



Conclusion

Another crucial factor is the precision of the search. If a system searches only for an exact word form rather than by lemma, certain grammatical forms may be excluded from the results. Conversely, a broader search might inadvertently include units that are spelled the same way but function differently. Therefore, after calculating frequencies, it is necessary to examine concordances, manually classify sample contexts, and explicitly state the search parameters in the article. Furthermore, the methods used for tokenization and grammatical tagging may differ between two corpora. Conclusions drawn without accounting for these discrepancies risk conflating linguistic features with technical specifications. To obtain reliable comparative results, it is necessary to specify the corpus name, the date of the search, the selected sub corpus, the search method (by lemma or word form), the total number of tokens, and the calculation formula used.

The National Corpus of the Uzbek Language and the National Corpus of the Russian Language provide the appropriate search tools for this study. However, conclusions regarding the actual frequency of lexemes and their predominant usage across specific text types must be drawn solely based on verified results derived from the corpora. Thus, by combining quantitative analysis with contextual interpretation, it becomes possible to conduct a scientifically grounded comparison of the lexical activity associated with the concept of “friendship” in both the Uzbek and Russian languages.

References:

1. McEnery, T., Hardie, A. Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press, 2012.
2. O‘zbek tili milliy korpusi. Qidiruv tizimi. https://uzbekcorpus.uz/enSearchLemma



3. Национальный корпус русского языка. Поиск по корпусу. https://ruscorpora.ru/en/search
4. Национальный корпус русского языка. Общая статистика корпусов. https://ruscorpora.ru/en/stats