Published: April 30, 2025 | Pages: 260-262

LINGUISTIC-STATISTICAL ANALYSIS OF WORD FORMATION IN LINGUISTICS

Unarova Dilafruz Abdimajit Qizi

Independent Researcher, Uzbek-Finnish Pedagogical Institute, Uzbekistan

ABSTRACT: In this article, the morphological structure of words in the Buryat language, especially the processes of suffix formation, is analyzed using linguistic and statistical methods using the literary texts of G.A. Dirkheeva. In the study, based on a large text corpus compiled from the works of Kh. Namsarayev, words were divided into morphemes, and their frequency, functional load, and variety of formation were studied. In addition, the stages of linguostatistical analysis of homonym grammatical forms were described and the methodology used in this direction was described. The article creates important theoretical and practical foundations for linguostatistical research.

KEYWORDS: Morphemes, suffix, linguistic statistics, word formation, homonym grammatical forms, automatic morphological analysis, corpus linguistics.

INTRODUCTION: In linguistics, the research conducted by G.A. Dirkheyeva is aimed at identifying the mechanisms of suffixal form formation in the Buryat language, the principles of dividing words into morphemes, and developing an algorithm for their automatic analysis based on literary texts. This article covers the main stages, methodology, and results of this analysis. The process of creating a corpus necessary for the linguostatistical study of homonym grammatical forms is also discussed. G.A. Dirkheyeva's article on the formation of suffixal forms in the Buryat language and the linguistic and statistical description of some features of word formation on the example of a literary text is devoted to the features of the structure of words in the Buryat language. The analysis is based on a reverse alphabetical-frequency dictionary compiled on the works of the representative of Buryat literature Kh. Namsarayev. The total volume of the processed text amounted to more than 272 thousand word usages, and the volume of the dictionary amounted to 36,540 word forms. For the purpose of automatic processing of the dictionary, the principles of dividing word forms into morphemes were developed, and the problems of dividing into morphemic components were considered. An automatic morphological analysis algorithm based on formally expressed elements of dividing into morphemic components was developed. A total of 5028 bases and 271 suffixes were extracted. The most productive and high-frequency bases and suffixes, as well as those that are rare and not described in the scientific literature, were identified. The functional load of individual suffixes, as well as the degree of diversity of a certain type of formation, were determined.

The list of articles of this kind can be continued with the works of L.V. Malakhovsky, O.M. Kim, and E.V. Chuyeva.

D. Orinbova's article on the basis of linguo-statistical analysis for lexical-semantic, linguo-stylistic and linguopoetic research is also noteworthy. In her opinion, linguo-statistical analysis ensures the accuracy of linguo-stylistic and linguopoetic analyses.

The analyses in these articles show that statistical analysis of grammatical forms can serve to reveal certain unstudied and undefined patterns in language.). Based on our observations, we summarize the stages of conducting a linguistic statistical analysis of homonyms as follows:

1. In the phase of identifying and selecting homonyms for analysis, the linguist identifies the homonymic grammatical form(s) he or she is interested in. For example, the homonymy of the possessive form of a noun and the homonymy of verb forms.

2. At the stage of creating or using a language corpus, a corpus of representative texts in the language under study is collected. The corpus can be general or specialized. For example, scientific literature, works of art, a corpus of colloquial speech, etc.

3. At the stage of automatic or manual tagging of the corpus, tagging is performed that allows you to unambiguously determine each word form and its grammatical affiliation in the corpus. This is very important for omoforms, since it is necessary to distinguish which grammatical form is expressed in each specific case. Often, special linguistic programs (parsers, taggers) are used for automatic tagging and subsequent manual verification.

4. At the stage of calculating the frequency of omoforms, the total number of uses of each omoform being analyzed in the corpus is calculated using a computer program.

5. At the stage of analyzing contextual distribution, it is studied in which sections of the corpus, between which words, in which syntactic constructions these or those omoforms occur. For this, corpus linguistics methods are used, for example, concordances that provide a list of word usages in context.

6. At the stage of statistical data processing, frequency and usage data are statistically analyzed to identify significant patterns and relationships. Various statistical tests and data visualization methods (graphs, diagrams) can be used.

7. At the stage of interpretation of the results, conclusions are drawn about the frequency, distribution of the studied homoforms and the factors influencing their use, based on statistical data and contextual analysis.

In this section, in order to carry out a linguostatistical analysis of homonymous grammatical forms, as mentioned in the previous sections, a corpus of the epics "Malikayi ayyor", "Kuntug'mish", and "Orzigul" was formed. The formation of the corpus for analysis is considered the most important of the 7 stages listed above, and below are the other steps necessary for the analysis.

REFERENCES

1. Avdina, A.I. Grammatical homonymy and homoforms - a scientific article in the field of linguistics and natural language processing.
2. Sobirov, M. Grammar of the Uzbek language. Tashkent, 2018.
3. Karimov, T. Fundamentals of linguostatistics. Tashkent, 2020.
4. Kholmurodov, N. Theory and practice of corpus linguistics. Tashkent, 2019.
5. Yusupova, D. Grammatical homonymy of the Uzbek language. Scientific article, 2021.
6. Mustafaev, R. Linguistics and natural language processing: theory and practice. Tashkent, 2022.