Horizon | Journal of Humanity and Artificial Intelligence

**Volume 3, Issue 3 | 2024** ***https://univerpubl.com/index.php/horizon***

**ISSN: 2835-3064**

*Article*

# Theoretical Problems of Taxonomy and Meronymics in The Uzbek National Corps

**Eshmuminov Askar Allamuradovich**

Head of the department, DSc of the Department of Uzbek Linguistic and literature, Termez State University, Termez City, Uzbekistan
eshmuminova@tersu.uz
tel: +998915812754

**Citation; Eshmuminov Askar AllamuradovichTheoreticalProblems of Taxonomy and Meronymics inThe Uzbek National Corps** Horizon: Journal of Humanity and Artificial Intelligence 2024, 3(3), 131-139

Received: 10th Oct 2024
Revised: 11th Nov 2024
Accepted: 24th Nov 2024
Published: 18th Dec 2024



**Abstract:** This article is present a case study of the monitor structure of the semantic search interface of the Uzbek language corpora, the structure of the combination of lexemes in the corpus, and the semantic tagging program projects were developed for the process of tagging the lexical units of the Uzbek language based on their mereological and taxonomic signs. The interface of the search window for lexical-semantic related units in the Uzbek language, the principles of classification based on thematic and ranking, and the linguistic model for language corpora have been determined. It is proved that taxonomically related lexemes in the explanatory dictionaries of the Uzbek language are quantitatively and thematically equal compared to meronymically related lexemes. Ensuring the active integration of the state language into modern information technologies and communications is one of the most important tasks facing the Uzbek corpus linguistics today.

**Keywords:** Corpus Linguistics, corpora, taxonomic, meronymic

## 1. Introduction

Corpus Linguistics is an independent branch of computer linguistics that deals with the development of principles for the creation and use of linguistic corpora (text corpus) using computer technology.[1] It began to develop as a separate branch of linguistics in the mid-1990s. There are some scientific notes about the corpus in science: to understand it as an electronic collection of texts equipped with a scientific apparatus[2]. as well as a reference to corpus linguistics, which is formed under the concept of "text corpus"[3] "Scope of electronic linguistic data for social and humanitarian research, with its own structure and designation"[4]. "The complex ontology of speech activity"[5] perception is observed. As a complex linguistic whole, the corpus not only retains a wide range of information about the content of a spoken/linguistic source, but also includes formal ways of presenting that information (such as

[1] Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – Санкт-Петербург, 2005. – 48 с.
[2] Махмараимова Ш.Т. Лингвокультурология // Ўқув қўлланма. – Тошкент: Чўлпон НМИУ, 2017. – Б. 31; Сабитова З. К. Новые лингвистические направления XX–XXI вв. www.ejournals.eu/pliki/art/3970/
[3] агель О.В. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении. http://sun.tsu.ru/mminfo/000349304/04/image/04-053.pdf
[4] Захаров В.П. Поисковые системы Интернета как инструмент лингвистических исследований // Русский язык в Интернете. – Казань, 2003. – С.52.
[5] Рыков В.В. Корпус текстов как новый тип словесного единства // Труды Междунар. семинара «Диалог-2003». – М.: Наука, 2003. С. 15–23.

word indexing, morphological information, etc.). Accordingly, the corpus can also be interpreted as a specially constructed semiotic system[6]

Corpus linguistics originated in Western Europe and the United States in the late 1960s as a discipline of text corpus analysis. The rapid development of computer and information and communication technologies led to the creation of corpus projects of various languages and sizes in the mid-1980s. During this period, the results achieved in corpus linguistics began to be widely applied to the language teaching process, and projects on corpus linguistics created in the world's leading higher education institutions became the topic of practice and the day. Through this, the historical, geographical, and social variations of the corpus approach and its optimal nature, reflecting its changing forms in the linguistic system, were revealed. This, in turn, provided an opportunity to learn the basic principles of corpus-based methods of linguistic analysis.[7]

The development of corpus linguistics and the creation of corpses is one of the most pressing issues for nations concerned about their future. In the first half of the twentieth century, it was possible to create a national corps only by hand. It took a long time and a lot of money. Therefore, the creation of text corpuscles is not so easy, and its use would be due to the interest shown by many. As mentioned above, with the development of information and communication technologies, the synchronization and systematization of materials related to the case has become easier, and costs have been reduced accordingly.

## 2. Materials and Methods

The methods of classification, description, comparison, and statistical analysis were used to illuminate the research topic in the establishment of the basis for the formation of the linguistic base of taxonomic and meronymic related lexemes for the corpora of the Uzbek language. It should be noted that higher education institutions around the world use linguistic corps to prepare control questions and lecture notes for students. Many students apply to the corps independently for independent work and project preparation. In this process, comparing students who use the corpus with those who do not use it, it can be concluded that a student who uses the corpus effectively learns the laws of language, as well as the features of the foreign language he learns more easily and quickly than a second category student. The linguistic corps is also an effective tool not only for scientific, but also for positively addressing educational issues. Researchers still classify corpus linguistics as a field in which theoretical foundations are now being formed.

The differences between traditional and corpus linguistics can be seen in the following places (Table 1.4).

Table 4.

Differences between traditional and corpus linguistics

| Features | In traditional linguistics | In Corpus Linguistics |
|---|---|---|
| Orientation | Language / language research | Speech research |
| Object of research | Explain speech phenomena through theory | Text corpus |

[6] Рыков В.В. Корпус текстов как новый тип словесного единства // Труды Междунар. семинара «Диалог-2003». – М.: Наука, 2003. – С. 21.

[7] Information and Communications Technology for Language Teachers. Введение в прикладное значение корпуса. Режим доступа: http://www.ict4lt.org/en/en_mod2-4.htm

| Methods | Qualitative methods | Quantitative methods |
|---|---|---|
| Attitude to the text | Abstraction | Accuracy |
| Focus on analysis | Focus on content | The focus is on shape |
| Methods of analysis | Limitations | Infinity |
| Types of analysis | Analyzes speech material based on intuition | Represents speech activity in the form of text |
| Methods of conveying information | Prefers logical thinking | Probability theory and statistics are used for the primary processing of speech material |
| The effectiveness of the analysis | The words in the text are analyzed separately | Works with linguistic information (word usage) in the text |
| Approach method | Deductive | Inductive |
| In relation to time | Diaxron | Synchronous |

**The essence of the corpus and its formation are aimed at:**

- be able to provide available information in text;

- Ability to provide as much information as possible depending on the size of the case;

- Ability to repeatedly use the data of the created corps to solve various problems, etc. To do this, you need to solve the following tasks facing the corps.

- it is necessary to define the principles that form the basis of the corpus, to determine what should be the standard corpus layout for different linguistic dimensions (this includes semantic, genre, methodological, semantic, morphological layout);

- to solve problems related to linguistics and literature with the help of the corps.

## 3. Results

This issue discusses the importance of taxonomic and meronymic lexemes in the corpus. Computer linguistics is a field that literally describes the operation of linguistic machines and the set of laws that govern them. The strength of computer linguistics lies in its ability to manage linguistic cues that are "saturated" with a particular set of meanings.[8] However, there is a direct "interaction" between computer and language; it also appears in the concept of computer-program-knowledge-language[9].

However, when this phenomenon is examined from the point of view of history and dialect, it shows that the clear signs of harmony of vowels are still within the law. The development of consonants is much more stable than that of vowels. This field has few differences within the Turkic languages. However, the interpretation of the Uzbek consonant system in the Altaic language family is determined by the fact that phonetic phenomena such as rotaicism and labdaism, which were the most important features of the consonants of that period, are preserved in the modern Uzbek literary language. While lexicon is the most rapidly evolving level of linguistics, onomastic units are considered to be the part that preserves linguistic relics brighter than other levels. Toponyms

[8] Шемякин Ю.И. Начала компьютерной лингвистики: Учеб. пособие. – М.: Изд-во МГОУ, А/О "Росвузнаука", 1992.
[9] Шемякин Ю.И. Начала компьютерной лингвистики: Учеб. пособие. – М.: Изд-во МГОУ, А/О "Росвузнаука", 1992.

are considered to be ancient onomastic units. However, they are also significant in that they retain the phonetic and morphological features of modern Uzbek literary language. Although the feature of modernity in anthroponyms prevails over other onomastic units, it can be seen that a careful study of this field has survived to the present day, embodying the semantics of desire in the names of Turkic-Uzbek people. The origin of zoonyms is associated with the earliest cases of the Turkic languages, and their emergence was confirmed by the fact that they relied on the mystery, character, color and other characteristics of living things. Relic situations at the level of morphology of the Uzbek language occur in the variation of morphological indicators. This is seen in the fact that forms relinquish their functional capabilities and move on to other tasks, or, more precisely, the fact that form-makers acquire the value of word-formation testifies to the preservation of the antiquity of language. Some of the affixes that acquired grammatical meaning in the ancient Turkic language have changed this value or have become rudimentary by moving to the core structure. The relictolinguistic approach to language proves that the archiforms that make up the grammatical number category are still in the grammatical function.

The basic concept of corpus linguistics is the linguistic corpus. Corpus texts are selected from the point of view of relevance to the problem, that is, from the thematic relevance that is important for linguists. The problem area is the thematic scale, which consists of two aspects, including texts on language and speech. While the language aspect is the most researched area, the speech aspect consists of contexts that reflect the language aspects. The problem of sorting texts by size is one of the most important issues in creating a corpus. This aspect ensures the perfection of the case. However, this aspect of the corpus contradicts its representativeness. Because the size of the text complicates the operation on the corpus, it means that the study of data on the corpus, depending on its size, can not provide a complete implementation. Thus, the goal of the corpus requires a narrow approach to its solution - to define the thematic area according to the scope of research. Such a choice should cover a particular criterion of the language under study in accordance with the events in the speech.

From the evolution of the World Corpus Linguistics, it is clear that almost all of the National Corps created so far will have a linguistic database interface for taxonomically related lexemes in the search menu. They are subordinated to the search engine in a certain order. Researcher A. Rakhimov explains: “The problem of modeling human thinking and language is a central problem in computer and corpus linguistics. In other words, research on artificial intelligence, natural language processing (NLP), linguistic processors, natural language interfaces for computers is carried out on the basis of creative cooperation in the fields of logic, grammar, semiotics, computer technology. The fact that neither the miraculous natural language nor the mysterious human thinking has yet been fully modeled is a promising direction in the science of computer linguistics.[10]” His ideas are the result of his own research.

An analysis of the interface of the linguistic database of lexemes with taxonomic and meronymic relations in the National Corpus of the Russian language showed that the units representing taxonomic and taxonomic and whole (meronymic) relations are arranged in a certain classification. Researcher U.Kholiyorov also said in his dissertation "Linguistic bases of the Uzbek language educational corps" about the location of the interface of the main menu of the Uzbek language National Corps: Particular attention should be

[10] Rahimov A. Kompyuter lingvistikasi asoslari – T.: Fan, 1953. –16 b.

paid to the features of the Uzbek language and its age-appropriate aspects "(see Figure 4.2).[11]

**4.Discussion**

The goals and objectives of the corpus are as follows: to develop a system that accommodates textual information that can be utilized several times to address different linguistic and literary problems. For this, it is presupposed to provide a clear framework of the defined corpus, which has to respond to principles and standards differentiated semantically, genre, methodology, and morphological dimensions. Particular attention should be paid to the taxonomic and meronomic lexis because these lexis require specific attention in corpora due to their direct connection to the structural relations that existed between elements of language and their applicability in various situations. Knowing these relations helps to advance the theory of media texts processing and analyzing them. And, of course, it is also important to consider historical and dia-tonic development of the language, which supports phonetic and morphological forms: rhotacism and labdaism and that are still present in the modern language, for instance, in Uzbek. This has some importance given to it so as to foster an understanding of the transformations going on the language and its structure. From the technical/n methodological viewpoints, a corpus implies the definitive structure of data necessary for the analysis of lexicon, morphology, and grammar. Some of them are still discussed in relation to sorting texts by the size and representativeness of the material since it is rather challenging to process a vast amount of data. A prime example of such work is A. Rakhimov's investigations of new opportunities for developing and implementing computer technologies and linguistic analysis especially, databases with taxonomical and meronymic lexis. This creates possibilities for the elaboration of detailer instruments, which serves as a benefit to corpus linguistics and further to Natural language processing.

**Literature review**

**Uzbek National Corps interface**

As we have seen, these classifications are further subdivided into internal taxonomic groups. As with human and animal taxonomy, plants can be studied in taxonomic groups. Professor A. Sobirov comments on the analysis of plants into taxonomic units. In modern taxonomy, plants are divided into the following units (taxa).

1. Section
2. Class.
3. Procedure.
4. Family.
5. Generation.
6. Tur.[12]

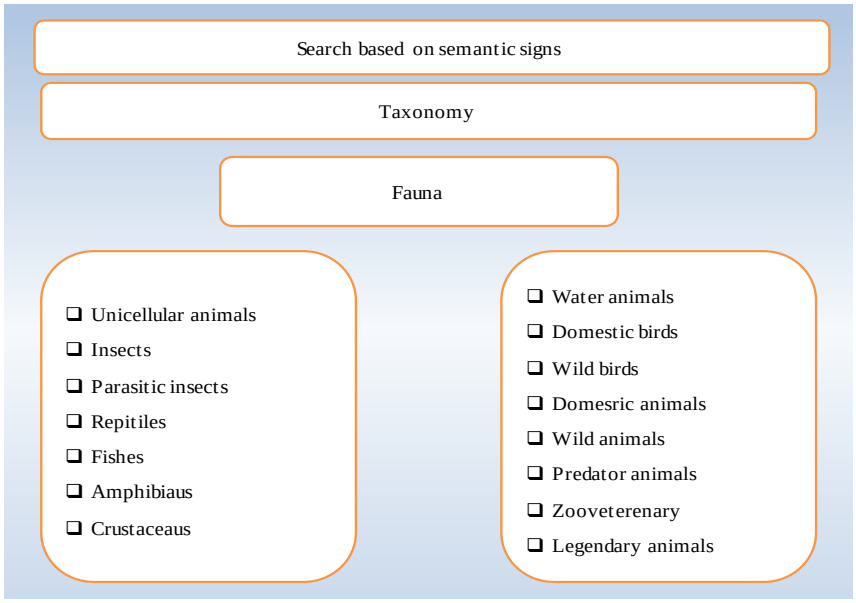
In addition to flora, the Uzbek fauna can be analyzed in terms of semantics; these taxonomic groups are again grouped internally. In line with these ideas, linguist A. Sobirov divides the animal world into the following taxonomic groups:

[11] Холиёров Ў. Ўзбек тили таълимий корпусини тузишнинг лингвистик асослари: Филол. фан. бўйича фалсафа доктори (PhD) дисс. – Термиз, 2021. –Б 38.

[12] Собиров А. Ўзбек тилининг лексик сатҳини системалар системаси тамойили асосида тадқиқ этиш. – Тошкент: Маънавият, 2004. – Б 93.

1. One-celled animals.
2. Insects.
3. Parasitic insects.
4. Reptiles.
5. Fish.
6. Amphibians.
7. Crustaceans.
8. Aquatic animals.
9. Domestic birds.
10. Wild birds.
11. Pets.
12. Live animals.
13. Predators.
14. Zooveterinary.
15. Legendary animals[13].

Based on the above classifications, a taxonomic search interface can be designed as follows:




The chapter entitled "Creating an Interface for a Linguistic Database of Meronymically Related Lexemes for Uzbek Language Corpus" discusses the problem of placing taxonomic units at the interface in the linguistic corpus. The linguistic base of meronymically related lexemes for Uzbek language corpora is based on research in Uzbek linguistics so far. In particular, linguist B. Kilichev analyzes the meronymic phenomenon in the following semantic areas:

1. Partonimi “Tana”
2. Partonim "Bosh"
3. "Foot" and "arm" sections
4. The "whole" whole
5. Particles of "tumor" semaphore

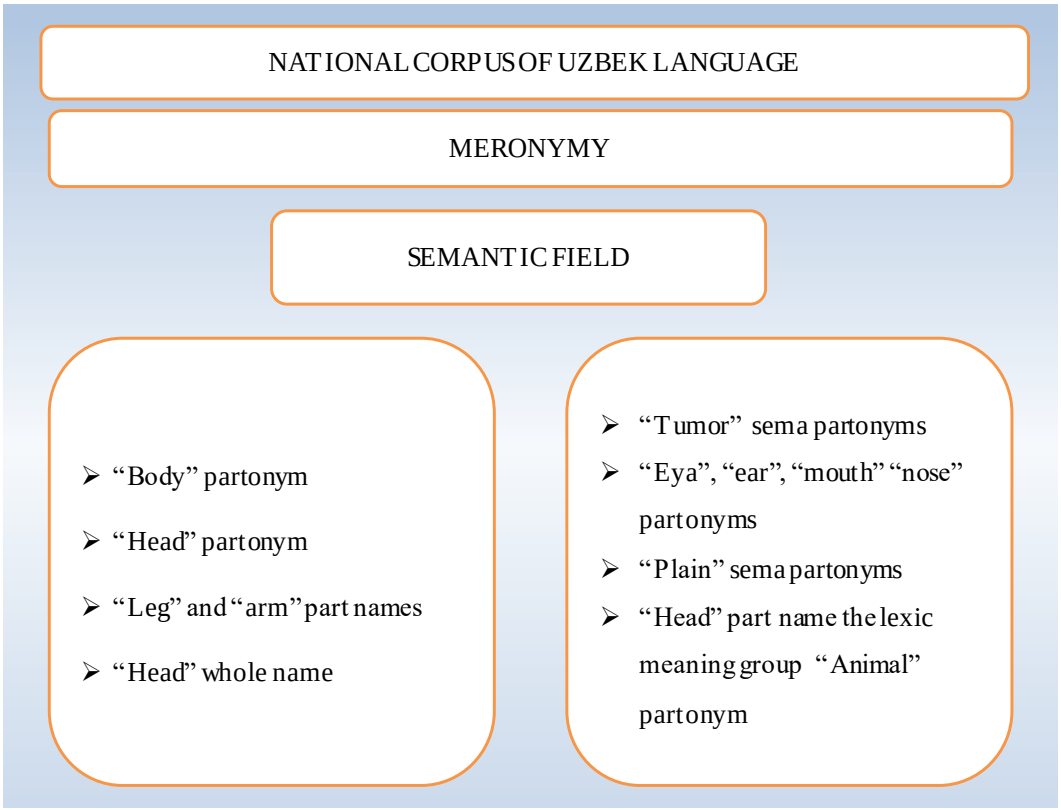
[13] Собиров А. Ўзбек тилининг лексик сатҳини системалар системаси тамойили асосида тадқиқ этиш. – Тошкент: Маънавият, 2004. – Б. 96.

6. Partonyms "eye", "ear", "mouth" and "nose"

7. Semantic partonyms "Plain"

8. The "head" section of the "animal" partonymous LMG[14]

This classification can be seen in the interface of the Uzbek National Corps as follows.




**Findings**

taxonomically and meronymically related lexemes in the Uzbek language were selected as study the general description of taxonomy and meronymy phenomena in linguistics and its importance in creating language corpora. Study the theoretical basis of analysis of language units based on the thematic area and the researching the appearance, content and functions of semantic search in the linguistic corpus interface.

## 1. Conclusion

The study of the Uzbek language's linguistic structure in the context of a diachronic and synchronous approach has led to the following conclusions:

As language develops in close connection with society, it moves away from its primitive boundaries. But no matter how far it goes from its oldest form, it is natural that it retains a certain amount of the features of the language of that period. Mankind is a subject that imagines the development of society, culture and science in combination with their history, and the study of the most ancient features of these units with the help of relictology can be seen in the study and interpretation of these remains in comparison with their current state. As an integral and most important part of the science of relictology, relictolinguistics is a branch of linguistics that studies the internal structure of languages formed and lost in the early twentieth century, the danger of

[14] Qilichev B. Leksikada so'zlararo butun-bo'lak munosabati. Monografiya. Toshkent. 2019. – 78.B.

forgetting, measures to ensure its survival. Linguistic relict, a unit of relictolinguistics, generalizes the earliest prescriptions of all levels of linguistics in diachronic and synchronous connections. This requires a distinction between the concepts of historical relic and linguistic relic. Since phonetics is one of the fastest changing fields of linguistics, it is difficult to imagine its earliest state. In particular, the fact that the Uzbek language has developed differently from other related languages makes it difficult to imagine its first features in the phonetic system. The vowels of the modern Uzbek literary language differ from other Turkic languages in that synharmonism is a rule, as they have an indifferent valueThe relictolinguistic study of the Uzbek language is based on the principle "good synchrony is not only static but also dynamic, on the contrary, good diachrony is not only dynamic but also static". can be a tool to help find solutions to existing problems.

**REFERENCES**


1. Zakharov, V. P. (2005). *Corpus linguistics: A study guide*. St. Petersburg.
2. Agel, O. V. (n.d.). *Corpus linguistics and its use in computerized language learning*. Retrieved from http://sun.tsu.ru/mminfo/000349304/04/image/04-053.pdf
3. Zakharov, V. P. (2003). Search engines on the Internet as a tool for linguistic research. In *Russian Language on the Internet* (p. 52). Kazan.
4. Nadim, M. H. (2021). Historical and etymological features of the ethnographic lexicon of Uzbeks in Northern Afghanistan. *Oriental Renaissance: Innovative, Educational, Natural and Social Sciences, 1*(2), 22-29.
5. Rykov, V. V. (2003). Text corpus as a new type of verbal unity. In *Proceedings of the International Seminar "Dialogue-2003"* (pp. 15–23). Moscow: Nauka.
6. Rykov, V. V. (2003). Text corpus as a new type of verbal unity. In *Proceedings of the International Seminar "Dialogue-2003"* (p. 21). Moscow: Nauka.
7. *Information and Communications Technology for Language Teachers*. (n.d.). Introduction to applied corpus linguistics. Retrieved from http://www.ict4lt.org/en/en_mod2-4.htm
8. Shemyakin, Yu. I. (1992). *Fundamentals of computer linguistics: A study guide*. Moscow: MGOU Publishing, Rosvuznauka.
9. Rahimov, A. (1953). *Fundamentals of computer linguistics*. Tashkent: Fan.
10. Sobirov, A. (2004). *Research on the lexical level of the Uzbek language based on the system of systems principle*. Tashkent: Ma'naviyat.
11. Qilichev, B. (2019). *The relationship between whole and part in the lexicon: A monograph*. Tashkent.
12. Nadim, M. H. (2021). Ethnocultural situation of Uzbek people in Northern Afghanistan. *Theoretical & Applied Science, 6*, 490-492.
13. Mengliyev, B. R., Hamroyeva, Sh., & Abdullayeva, O. (2023). Scopus-based bibliometric analysis on corpus linguistics for the period of 2017-2021. *E3S Web of Conferences, 413*, 03008.

14. Crosthwaite, P., Ningrum, S., & Schweinberger, M. (2023). Research trends in corpus linguistics: A bibliometric analysis of two decades of Scopus-indexed corpus linguistics research in arts and humanities. *International Journal of Corpus Linguistics, 28*(3), 344-377.
15. Park, H. (2018). A bibliometric analysis of corpus linguistics: 1997–2016. *Corpora, 13*(1), 83-104.
16. Liao, S., & Lei, L. (2017). Bibliometric analysis of corpus research in China: 2000–2015. *Corpora, 12*(1), 93-120.
17. Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.
18. McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
19. Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.
20. Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.