INTERNATIONAL MULTI DISCIPLINARY JOURNAL FOR RESEARCH & DEVELOPMENT
eISSN 2394-6334 Impact factor: 8.854 Volume 13, issue 03 (2026)

PRINCIPLES OF LINGUISTIC ANNOTATION IN LANGUAGE CORPORA

Abdurahmanova Sayyora Bozorboy kizi
PhD, teacher at TSUI
sayyoraabdurahman@gmail.com, 909985059

Annotatsiya: Mazkur maqolada korpus lingvistikasi, xususan, dialektal korpuslarning lingvistik annotatsiyasi masalalari yoritilgan. Annotatsiya turlari — leksik, morfologik, sintaktik, fonetik, semantik, pragmatik va diskursiv — batafsil tahlil qilinib, ularning har biri dialektal farqlarni aniqlashdagi roli misollar bilan koʻrsatib berilgan. Shuningdek, korpus annotatsiyasining tarixi, avtomatik va qoʻlda tuzatiladigan annotatsiyalar, annotatsiyalash vositalari va formatlari haqida ham ma'lumotlar beriladi. Dialektal birliklarning tahlili orqali Oʻzbekiston hududlaridagi shevalar oʻrtasidagi fonetik, semantik va pragmatik tafovutlar yoritiladi. Maqola korpus lingvistikasida dialektologik tadqiqotlarni tizimli asosda olib borish uchun muhim nazariy va amaliy asoslarni taqdim etadi.

Kalit soʻzlar: Korpus lingvistikasi, dialektal korpus, annotatsiya, leksik annotatsiya, morfologik annotatsiya, sintaktik annotatsiya, semantik annotatsiya, fonetik annotatsiya, pragmatik annotatsiya, diskursiv annotatsiya, teglash tizimi, XML, TEI, Universal Dependencies, avtomatik annotatsiya, qoʻlda tuzatish, nutq qismlari, lingvistik tahlil, shevalar farqi.

Annotation. This article discusses corpus linguistics with a particular focus on the linguistic annotation of dialectal corpora. It provides a detailed overview of various annotation types — lexical, morphological, syntactic, phonetic, semantic, pragmatic, and discursive — and explains their roles in identifying dialectal differences, supported by relevant examples. The paper also outlines the history of corpus annotation, including automatic and manually corrected methods, and presents commonly used annotation tools and formats. Through the analysis of dialectal units, the article highlights phonetic, semantic, and pragmatic variations across different regional dialects in Uzbekistan. It offers both theoretical and practical foundations for conducting systematic dialectological research within the framework of corpus linguistics.

Keywords Corpus linguistics, dialectal corpus, annotation, lexical annotation, morphological annotation, syntactic annotation, semantic annotation, phonetic annotation, pragmatic annotation, discursive annotation, tagging system, XML, TEI, Universal Dependencies, automatic annotation, manual correction, parts of speech, linguistic analysis, dialectal variation.

Corpora have become important scientific tools for researchers in natural language processing, speech recognition, spoken language materials, as well as in theoretical linguistics. A corpus is not merely a collection of simple texts, but a body of electronic data in which each linguistic unit is annotated with linguistic and extralinguistic information. In specialized corpora, especially those based on natural language materials, additional information about texts, authors, and the contexts of data collection is of great importance. This information is used in comparative studies, as the texts selected for comparison are based on naturally occurring data.

In the annotation of linguistic units using corpora, special tags are employed, for example, to indicate the class of a word or its grammatical category. Annotated corpora include various types of annotation: part-of-speech tagging, as well as prosodic, semantic, anaphoric, and discourse annotations. Grammatically annotated corpora represent the most common type of annotated corpus. In such corpora, linguistic units are grouped into categories and their grammatical features are described.

Looking at the history of corpus annotation, the first annotations were carried out in the Brown Corpus by Francis and Kucera. In this corpus, parts of speech were tagged, and soon additional linguistic tags were introduced. A total of 77 different tags were used. Subsequently, the LOB Corpus experimented with verifying annotations using probability theory. During this process, tagging programs such as CLAWS were developed, but due to high error rates, manual correction was required. After the 1990s, specialized programs began to be developed for other languages. Grammatical tagging gained momentum after 1987. This shows that as soon as corpora began to be created, the processing of texts and natural language processing issues came to the forefront.

Corpus annotation is the practice of adding linguistic information to the electronic form of a text. Some sources distinguish three types of corpus annotation: structural markup, part-of-speech tagging, and grammatical annotation. Structural markup includes descriptive information about texts, such as titles, authors, and other contextual data. In part-of-speech tagging, each word is annotated with tags indicating its grammatical class. Linguistic annotation in corpora is often carried out in formats such as XML, where words and phrases are marked with special codes.

There are three main approaches to linguistic annotation: automatic annotation, semi-automatic annotation, and manual correction. Automatic annotation systems, such as those developed by Garside and Smith, provide high accuracy but still contain errors. Manually corrected annotations are performed by humans, but even these are not guaranteed to be perfect.

Morphosyntactic annotation is the most common type of annotation in corpora, involving the classification of each word according to its grammatical category. Semantic annotation focuses on identifying the lexical and semantic groupings of words. Discourse annotation describes anaphoric relationships and other discourse elements. Prosodic annotation provides transcription of intonation, stress, and pauses, which is particularly important in the analysis of spoken language.

It is important to adhere to the principles of linguistic annotation. These principles include proper documentation of the annotation process, awareness of possible errors, and ensuring that systems are understandable. Annotations in corpora should be developed based on uniform standards, as corpora in different languages may have their own specific features.

Thus, corpus annotation is an essential tool in linguistic research, natural language processing, and other areas of linguistics, as it provides the necessary information to analyze and understand the linguistic properties of texts. Linguistic annotation of a dialectal corpus is the process of systematically marking and analyzing language units, which is important for studying different dialects, building structured databases, and conducting linguistic analysis.

Annotation involves identifying and clearly marking grammatical, lexical, syntactic, phonetic, and other linguistic aspects of language units. The process of annotating a dialectal corpus includes the following main components and stages:

Lexical annotation focuses on the meanings of words and their usage across different dialects. At this stage, dialect-specific lexical units, synonyms, antonyms, and dialectal words are annotated. The semantic properties of words are also considered, as the same word may have different meanings in different regions. For example, in the Tashkent dialect, the word “yaxshi” means “good,” but in other regions it may carry different semantic nuances.

Morphological annotation deals with the structure and morphological features of words. In a dialectal corpus, words are classified according to their morphological forms (nouns, verbs, adjectives, etc.), and dialectal differences in grammatical forms are indicated. For example, the verb “boraman” in the Tashkent dialect may appear as “boramiz” in the Fergana dialect.

Syntactic annotation involves the analysis of sentence structure and word combinations. Different dialects may have different word orders and syntactic constructions. For example, the sentence “Men kitobni o‘qiyapman” in the Tashkent dialect may appear as “Kitapdi ben o‘qiyappan” in another dialect.

Phonetic annotation identifies pronunciation differences and phonetic features. It captures how sounds vary across dialects. For instance, in some dialects, the vowel “a” may shift toward “o.”

Pragmatic annotation reflects the social and cultural context of word and phrase usage. Certain expressions may be specific to particular social groups or situations. For example, the word “qilov” may have different meanings in different regions.

Semantic annotation examines changes in meaning across dialects. A word may carry different meanings or extended meanings in different regions. For instance, “yaxshi” may mean “good,” “decent,” or “healthy” depending on the context.

Multidimensional annotation involves combining several linguistic layers—phonetic, morphological, syntactic, semantic, and pragmatic—to show their interrelationships and highlight the unique features of each dialect.

Discourse annotation analyzes contextual and pragmatic features of texts, including communicative purpose and structure. In dialectal corpora, it helps identify differences between dialects and their social and cultural contexts. For example, a sentence like “Men hozir o‘qivomman” indicates an ongoing action and reflects specific pragmatic usage in the Tashkent dialect.

In discourse annotation within dialectal corpora:

1. Contextual analysis shows how dialect-specific lexical and grammatical units function in social contexts.

2. Pragmatic analysis reveals how intentions such as offering help or refusal vary across dialects.

Tagging systems are used to label linguistic units. Each word or phrase is assigned specific tags indicating grammatical, lexical, phonetic, or syntactic features (e.g., “NN” for noun, “VB” for verb, “JJ” for adjective).

Annotation tools and formats include software and frameworks such as XML, TEI, UIMA, and Universal Dependencies, which allow systematic annotation and analysis of linguistic data.

The process of annotating dialectal corpora enables a clear and systematic representation of linguistic features, including semantic, morphological, syntactic, phonetic, and pragmatic aspects. It facilitates deeper study of dialects and helps identify and analyze dialectal differences.

Each annotated unit reflects phonetic, morphological, syntactic, and semantic differences across dialects. Annotation is the process of adding explanatory information to words or linguistic units in a corpus. Linguistic annotation provides additional interpretation of each unit from a linguistic perspective.

Today, corpus linguistics encompasses these annotation processes as an essential component. A corpus is not only a collection of texts but also an electronic database that provides structured and reliable theoretical data for scientific research. Therefore, national corpora should be widely used not only by learners of the Uzbek language but also by linguists and researchers. In this process, linguistic annotations play a crucial role. Annotation can be carried out automatically, semi-automatically, or manually.

FOYDALANILGAN ADABIYOTLAR

1. Garside R., Leech G., McEnery T. Corpus annotation. – Routledge, 1997. – P. 292.

2. Abdullayeva O. O‘zbek tilining internet axborot matnlari korpusini shakllantirishning nazariya va amaliy asoslari. Filol. fan. fals. dok. (PhD) …diss.– Toshkent, 2022. – B. 102.

3. Abdullayeva O. O‘zbek tilining internet axborot matnlari korpusini shakllantirishning nazariya va amaliy asoslari. Filol. fan. fals. dok. (PhD) …diss.– Toshkent, 2022. – B. 105.

4. McEnery T., Hardie A. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press, 2012. –P.30.

5. Vakhobova, M. (2022). Main principles of ICT-assisted language learning and teaching. Архив научных исследований, 4(1).

6. Vakhobova, M. A. (2022). INNOVATIVE METHODS OF THE DISTANCE LEARNING PROCESS IN MODERN UNIVERSITIES. Oriental renaissance: Innovative, educational, natural and social sciences, 2(5-2), 641-645.

7. Vakhobova, M. (2022). The Richness of the English Language. Архив научных исследований, 4(1).

8. Vakhobova, M. (2022). INNOVATIONS IN EDUCATION AS A NECESSARY CONDITION FOR THE DEVELOPMENT OF CREATIVITY OF UNIVERSITY STUDENTS. Архив научных исследований, 4(1). Vakhobova, M. (2022). THE GENERAL CHARACTERISTICS OF TEACHING AND READING COMPREHENSION. Архив научных исследований, 4(1). Vakhobova, M. (2022). THE USE OF GAMES AS A STRATEGY OF DEVELOPING COMMUNICATIVE COMPETENCE OF LEARNERS. Архив научных исследований, 4(1).

9. Abdurahmanova, S. B. K. (2023). Basics of composition of the corpus of lacunar units in uzbek dialects. Oriental renaissance: Innovative, educational, natural and social sciences, 3(6), 1150-1153.

10. Abdurahmanova, S. B. Q. (2026). O ‘ZBEK TILIDA YARATILGAN KORPUSLAR TAVSIFI. Oriental renaissance: Innovative, educational, natural and social sciences, 6(2), 10-15.

11. ABDURAHMANOVA, S. (2024). SHEVALAR KORPUSINI YARATISHNING NAZARIY ASOSLARI (O ‘ZBEK KORPUS LINGVISTIKASINING SHAKLLANISHI VA TARAQQIYOTI). «ACTA NUUz», 1(1.10. 1), 273-275.

12. Abdurahmonova, S. S. (2022). FACTORS FOR FORMATION OF PEDAGOGICAL CULTURE OF PRESCHOOL EDUCATIONAL TEACHERS. Oriental renaissance: Innovative, educational, natural and social sciences, 2(6), 170-173.

13. https://www.philol.msu.ru/~ref/2014/2014_GorinaOG_diss_13.00.02.pdf

14. https://escholarship.org/uc/item/09v5z6fg

15. http://www.natcorp.ox.ac.uk/

16. http://www.natcorp.ox.ac.uk/ ; https://cldf.clld.org/

17. https://www.ice-corpora.uzh.ch/en.html

18. https://aclanthology.org/L14-1287/

19. Abdurahmonova N. Semantik annotatsiyalangan korpus yaratish tajribasidan “O‘zbek tilining milliy korpusi: muammolar va vazifalar” mavzusidagi xalqaro ilmiy-amaliy anjumani ma’ruzalar to‘plami. –Samarqand, 2023.

20. Гулямова Ш. Ўзбек тили семантик анализаторининг лингвистик асослари. Филол. фан. док. (DSc) ... дисс. автореф. – Фарғона, 2022.

21. BAKHTIYOROVNA, S. D. (2026). INDIVIDUALIZING INDEPENDENT LEARNING WITH THE USE OF ARTIFICIAL INTELLIGENCE. Shokh Articles Library, 1(1).

22. Baxtiyorovna, S. D. (2025). OLIY TA'LIMDA MULTIMEDIYANI QO ‘LLASH TAJRIBASI. ZAMONAVIY TA'LIMDA FAN VA INNOVATSION TADQIQOTLAR, 3(10), 283-286.

23. Bakhtiyorovna, S. D. (2025). THE IMPORTANCE OF MULTIMEDIA IN ORGANIZING STUDENTS'INDEPENDENT WORK. INTERNATIONAL JOURNAL OF

ADVANCED RESEARCH IN EDUCATION, TECHNOLOGY AND MANAGEMENT, 4(3), 153-161.

24. Sulaymanova, D. B. (2024). Texnika yo’nalishlarida dasturlashning kasbiy afzalliklari. Zamonaviy ta'limda fan va innovatsion tadqiqotlar jurnali, 203-211.

25. Sulaymanova, D., Abduganieva, Y., & Miratoev, Z. (2023). Modeling roll contact curves of a squeezing machine. In E3S Web of Conferences (Vol. 443, p. 03006). EDP Sciences.