



Dialectology and Corpus Linguistics: World Experience

Farzona Murodova

Termez State University,

Philology and Language Teaching: Uzbek Language Department

Phone: +998 94 015 12 06

E-mail: fmurodova906@gmail.com

Abstract: This article is devoted to studying the world experience in the fields of dialectology and corpus linguistics. Dialectology is a scientific field that studies the regional and social variants of a language, while corpus linguistics focuses on analyzing large collections of texts — corpora — with the help of computers. The article discusses advanced research conducted around the world, major dialect and language corpora, their practical application in linguistics, as well as the influence of modern corpus methodologies on dialectology. At the same time, the article demonstrates the possibilities of studying international experience on the dialects of the Uzbek language and analyzing them on a corpus basis.

Keywords: dialectology, corpus linguistics, dialect, vernacular, language corpus, accent/sub-dialect, linguistics, world experience, research.

Dialectology and corpus linguistics, as new research directions in the field of linguistics, are of great importance. While dialectology is aimed at studying the regional and social variants of a language, corpus linguistics makes it possible to systematically analyze language material with the help of computers. In recent years, in world linguistic practice, the integration between these two directions has been strengthening, helping to improve the quality of scientific research and to expand language resources.

Dialectology comes from the Greek word “*dialektos*” and studies the existing dialects and sub-dialects of a given language. “*Dialektos*” means “*dialect/vernacular*,” and “*logos*” means “*doctrine*,” that is, it literally means “*the doctrine of dialect*.” Almost all languages existing in the world consist of their own particular dialects or internal varieties. Although each of the Turkic languages consists of several dialects and sub-dialects, the Uzbek language stands out among them for its especially rich dialectal diversity. One of the experts in dialectology, Professor E.D. Polivanov, stated that none of the Turkic languages differs between its dialects as sharply as the Uzbek language does, and that this naturally makes the task of choosing a single dialect as the basis for the literary language extremely difficult.

Dialectology can be divided into two types: 1) descriptive dialectology, or dialectography; 2) historical dialectology. Descriptive dialectology, or dialectography, records and writes down the phonetic, lexical, and morphological features specific to local dialects and sub-dialects, and maps them onto a linguistic map. Historical dialectology, alongside the dialectal features of a language, studies the origin of these features, their development, their changes across different periods, their



relationship with kindred languages, and the role of both related and unrelated languages in the formation of these dialects.

The subject matter studied by dialectology includes the terms dialect, sub-dialect (sheva), and dialect group (lahja). The word “dialect” (originally meaning “tribal language”) unites sub-dialects whose linguistic features are largely similar. In linguistics, it is often used in the sense of “lahja” (dialect group) and sometimes in the sense of “sheva” (sub-dialect). In scholarly works created by scholars, the terms sheva, dialect, and lahja are distinguished, although cases of mixing them up also occur. “Dialect” is an international term and should be used in the general linguistic sense of “dialect.”

Sheva (sub-dialect) comes from the Persian words meaning “manner,” “style,” “way,” and in scholarship refers to a part of a language used in a small area, distinguished by its own particular lexical, phonetic, and grammatical features. (For example, the Toshloq sub-dialect, the Qoranoq sub-dialect, the Turk sub-dialect, the Qipchoq sub-dialect, and others.)

Lahja (dialect group) comes from an Arabic word meaning “manner,” “style.” As a dialectological term, it denotes the largest grouping of sub-dialects and dialects. As noted in A.N. Qayumov’s work “Uzbek Dialectology” (1992), lahja is a form of folk speech, used in a particular area, that differs from the literary language in its phonetic, grammatical, and lexical features.⁶ In A. Madrahimov’s book “Dialectology of the Uzbek Language” (2007), lahja is described as a branch standing at a level higher than sub-dialects, generalizing a group of several closely related sub-dialects.⁷ Thus, lahja is a linguistic unit that brings together local features in relation to the literary language, and the smallest unit composing it is the sub-dialect (sheva). That is, there exists a hierarchy in the form: sub-dialect (sheva) → dialect group (lahja) → literary language. In the explanatory dictionary of the Uzbek language, lahja is explained as “a form of a language that has taken shape in a particular territory and possesses its own particularities in pronunciation and vocabulary.” These definitions emphasize three elements as the main features of lahja — differences in phonetics, lexicon, and grammar. For example, in the Tashkent dialect group, the literary form “borayapman” (“I am going”) is pronounced as “borvoman,” and “ketayapman” (“I am leaving”) as “ketvoman” — an example of phonetic simplification. Lexical differences also occur: the literary word “bolalar” (“children”) is pronounced as “bachalar” in the Tashkent dialect group.

The importance of compiling dialectological corpora lies in the fact that, in economically disadvantaged regions, local sub-dialects spoken by only a few tens of thousands of people are gradually disappearing — recording and preserving them in digital corpora is today an urgent task. At present, the world’s population speaks close to 7,000 languages, and each of these languages is divided into several dialects according to social and territorial factors.



Corpus and corpus linguistics took shape as an independent scholarly direction in the 1960s.

The term “corpus” is polysemous and is used in linguistics to denote various objects; in general, it is a collection of authentic texts belonging to a particular language. According to K. Bausch’s definition: “A corpus is a finite collection of language material consisting only of (written or spoken) texts or parts of them.” C. Meyer, in turn, defines a corpus as a collection of authentic spoken or written texts, compiled in a balanced and stratified manner, aimed at representing the diversity of a language. Corpus linguistics shows that analyzing language yields more accurate results when authentic data, collected under natural conditions without experimental interference, is used. Large volumes of text give linguists the opportunity to carry out quantitative analysis and to identify linguistic phenomena that are difficult to study qualitatively.

The text-corpus method uses a collection of texts in natural language to identify the abstract rules that govern the language. Corpora have been used not only for linguistic research but also in compiling dictionaries and reference grammars. This practice began in 1969 with the dictionary “The American Heritage Dictionary of the English Language” and continued with the publication of the grammar reference “A Comprehensive Grammar of the English Language” in 1985. Scholars such as B. Yelov and N. Abdurahmonova have taken important steps toward digitizing the Uzbek language into corpora, drawing on world corpus experience while proposing solutions suited to the features of the national language.

In E. Finegan’s textbook, a corpus is defined as a collection of representative texts, in machine-readable format, containing information about the speaker, author, addressee, or audience. According to the definition by T. McEnery and A. Wilson, a corpus is a collection of pieces of language selected according to explicit criteria for use as a sample of the language. J. Sinclair — the founder of corpus linguistics — defines a corpus as “a collection of pieces of electronic text selected according to external criteria to represent, as fully as possible, the language as a whole.” In the Russian National Corpus, a corpus is defined as “an information-reference system based on a collection of texts in electronic form in a given language.” These definitions reveal four common features of a corpus: it is purposefully created, in electronic format, representative, and contains metalinguistic information. E. Finegan describes corpus linguistics as a scholarly activity aimed at studying the natural use of language, based on the compiling and use of corpora.

Among the largest examples of dialectal corpora in world linguistics, the Helsinki Dialect Corpus of British English stands out. It brings together phonetically transcribed audio recordings collected from the speech of residents of East Anglia, Southwest England, and Lancashire. In creating a corpus of dialects, the initial and most important task is to mark up (annotate) the texts phonetically, lexically, and grammatically. Marked-up (annotated) and unmarked (unannotated) corpora are distinguished. Markup (annotation) is the process of attaching special tags to a text and its components. Tags are of two kinds: (1) linguistic tags — describing the lexical, grammatical,



and other features of text elements; (2) extralinguistic tags — representing external information such as the author, year and place of publication, genre, and theme.

Annotation is carried out in three directions: (1) the observation stage — recording the place where the text was collected, the ethnic composition of the population, and the respondent's profession and level of education; (2) phonetic annotation — tagging the front/back articulation of vowels, vowel harmony, stressed/unstressed features, and the plosive–fricative, voiced–voiceless categories of consonants; (3) dialectal lexical meta-annotation — linking each dialectal word to its literary-language equivalent, with synonyms, homonyms, antonyms, and paronyms marked separately. Carrying out lexical-semantic tagging of words across different dialects is a particularly complex process. An annotated corpus retains linguistic information within itself; Professor Richard Xiao of Lancaster University has discussed in detail how such a corpus differs from a plain text. An unannotated corpus, by contrast, is unprocessed, “raw” text that does not carry linguistic information.

World experience shows that a corpus is never a finished project — it is constantly being updated. Just as, alongside general corpora in English, specialized corpora have emerged (for social media, medicine, law), specialized corpora for various fields are likewise taking shape in the Uzbek language alongside the national corpus. In this respect, the Uzbek language corpus project is also following world experience.

In conclusion, the integration of corpus linguistics and dialectology opens up new prospects in linguistics. Through corpora, regional, social, and lexical differences are depicted precisely — this provides faster and more scientifically reliable results compared to traditional field research. Modern software tools make it possible to systematically collect, store, and analyze dialectal material. From the standpoint of Uzbek linguistics, creating corpora of national sub-dialects and dialect groups serves not only to preserve and develop the language but also increases opportunities for cooperation with the international scholarly community. For this reason, developing the integration of dialectology and corpus linguistics, and applying advanced world experience in national research, is of not only scholarly but also practical significance.

References:

1. To'ychiboyev B., Xasanov B. *Uzbek Dialectology*. – Tashkent, 2004. – p. 4.
2. Ashirboyev S. *Uzbek Dialectology*. – Tashkent, 2016.
3. Qayumov A.N. *Uzbek Dialectology*. – Tashkent, 1992.
7. Madrahimov A. *Dialectology of the Uzbek Language*. – Tashkent, 2007.
8. Ashirboyev S., Turdialiyeu A. *Software for Incorporating the Lexicon of Uzbek Dialects into a Corpus*. – p. 25.
9. Xidirov O. *Corpus and Corpus Linguistics*. – p. 1.



10. Meyer Ch.F. *English Corpus Linguistics*. 2nd ed. – Cambridge University Press, 2023. – p. 4.
11. Hunston S. *Corpus Linguistics // Encyclopedia of Language and Linguistics*. 2nd ed. – Oxford: Elsevier, 2006. – pp. 234–248.
12. Sinclair J. *The Automatic Analysis of Corpora // Svartvik J. (ed.) Directions in Corpus Linguistics*. – Berlin: Mouton de Gruyter, 1992.
13. Abjalova M. *Corpus Linguistics*. – Tashkent, 2022. – p. 7.
14. Primova M. *The Structure of the National Corpus of the Uzbek Language: Main Components and Types of Information*. – p. 219.
15. *Computer Linguistics: Problems, Solutions, Prospects: International Scientific-Practical Conference*. – p. 182.