Nozimjon ATABOYEV
student at masters department
Uzbekistan State World Languages University
nozimjon1992.na@gmail.com

THE MAIN THEORETICAL PROBLEMS OF CORPUS LINGUISTICS

Ушбу мақола бугунги кунда ривожланиб бораётган корпус лингвистикасининг илмий муаммоларини ўрганишга бағишланган. Лингвистлар томонидан ҳали ечими топилмаган ушбу муаммолардан бир нечтаси санаб ўтилган ва мавжуд мулоҳазалар нуқтаи назаридан муҳокама этилган.

В настоящей статье рассматриваются основные предпосылки и проблемы корпусной лингвистики. В статье приводятся мнения различных ученых об указанных темах.

The present article discusses main assumptions and problems of Corpus linguistics. In the article the mentioned problems are discussed from the point of views of several linguists.

Калит сўзлар: корпус лингвистикаси, корпус, фан, метод, методология, парадигма, салбий томонлари, классификациялаш.
Ключевые слова: корпусная лингвистика, корпус, предмет, метод, методология, парадигма, негативные стороны, классификация.
Key words: corpus linguistics, corpus, science, method, methodology, paradigm, negative sides, classification.

Corpus Linguistics is an approach that aims at investigating language and all its properties by analyzing large collections of text samples. At present, Corpus linguistics is becoming more and more applied in different linguistic spheres.

As is known, “A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research”(1, 16).

As the observation of theoretical literature has shown, there are several problems of Corpus linguistics. They are as follows:

- metalanguage of Corpus linguistics;

- the size of corpora;
- the classification of corpora;
- the problematic features of Corpus linguistics;
- the problem of authenticity of the texts in corpora;
- the measurement of the corpora database;

Let's discuss some of them in detail. First of foremost problem concerns metalanguage or terminology.

According to I.F. Ganieva, there are two different *terms* that can be used with no difference in the field. They are '*language corpus'* and *'linguistic corpus'*. She mentions A.A. Polikarpov's point of view in this challenge. According to Polikarpov, linguistic corpus is not appropriate term to use as a synonym of language corpus, because linguistic corpus is a corpus about the language, theories on language, but not of the language(3, 105).

The following problem to be discussed is the question as "What is corpus linguistics?" – is it a discipline, a methodology, a paradigm or none or all of these? This question does not have any definitive answers. However, at the same time, there are several options that can be observed via variety of definitions and descriptions of corpus linguistics by different linguistic scholars. In fact, Aarts, one of the founding fathers, and Meijs, as the first book dedicated to the subject, is often identified as the source of the term c*orpus linguistics*, although the term had in fact been used previously, for example, in Aarts and van den Heuvel. From that time this raises one of the recurrent concerns over talking about *corpus linguistics*, and may account for the preference for alternatives. In terms of what corpus linguistics 'is', not only have various definitions been offered, but alternatives have been explicitly addressed and rejected. These include, as we shall see: *corpus linguistics* is a *tool*, a *method*, a *methodology*, a *methodological approach*, a *discipline*, a *theory*, a *theoretical approach*, a *paradigm* (theoretical or methodological), or a combination of these.

Corpus linguistics is a discipline:

— Aarts is reported as commenting that the term was coined with some hesitation "because we thought that it was not a very good name: it is an odd *discipline* that is called by the name of its major research tool and data source. Perhaps the term has outlived its usefulness by now";

— Stubbs describes linguistics as an "applied social *science*"(7, 3);

— Teubert states that "linguistics is not a *science* like the natural sciences whose remit is the search for 'truth'. It belongs to the humanities, and as such it is a part of the endeavour to make sense of the human condition"(4, 7).

Corpus linguistics is a paradigm:

— The notion of corpus linguistics as a paradigm is taken up by Gries, but the methodological conceptualisation is favoured, as he states that "over the past few decades, corpus linguistics has become a major methodological paradigm in applied and theoretical linguistics"(3, 191)

Corpus linguistics is an approach:

— Leech argued that "computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject" and goes on to describe the characteristics of computer corpus linguistics as a new paradigm(11, 106).

— Teubert also emphasises the theoretical conceptualisation and describes corpus linguistics as "a theoretical approach to the study of language"(4, 2).

— Mahlberg describes corpus linguistics as "an approach to the description of English with its own theoretical framework", and to emphasise this employs the term "corpus theoretical approach"(12, 363).

Corpus linguistics is not a method:

— it is on the claim to scientific method that Chomsky criticized corpus linguistics, stating "my judgment, if you like, is that we learn more about language by following the standard method of the sciences. The standard method of the sciences is not to accumulate huge masses of unanalyzed data and to try to draw some generalization from them"(13, 97).

Corpus linguistics is a methodology:

— Stubbs rejects the limited definition of corpus linguistics as a methodology, and, commenting on Sinclair 1991, he notes that "in this vision of the subject, a corpus is not merely a tool of linguistic analysis but an important concept in linguistic theory"(5, 23–24).

— Tognini-Bonelli described corpus linguistics as a "pre-application methodology" which possesses "theoretical status"(4, 1). In directly addressing the issue she sees the difference of perception as stemming from the type of corpus linguistics which the researcher practices: "there is still disagreement on whether corpus linguistics is mainly a methodology or needs its own theoretical framework. Advocates of corpus-driven approaches to the description of English claim that new descriptive tools are needed to account for the situation of real text, and ideas of theoretical frameworks to accommodate such tools have started to emerge".

— Thompson and Hunston state that "at its most basic corpus linguistics is a methodology that can be aligned to any theoretical approach to language"(5, 8).

— McEnery, Xiao and Tono, note that as "corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning, it certainly has a theoretical status. Yet theoretical status is not

theory in itself"(6, 7–8); they therefore conclude that corpus linguistics is a methodology.

— Corpus linguistics is also defined as a methodology in McEnery and Wilson and Meyer, and as "an approach or a methodology for studying language use" in Bowker and Pearson.

— McEnery and Gabrielotos note that "corpus linguistics may be viewed as a methodology, but the methodological practices adopted by corpus linguists are not uniform"(7, 44),

— Teubert also comments on the diversity of methods, and states that "corpus linguistics is not in itself a method: many different methods are used in processing and analyzing corpus data. It is rather an insistence on working only with real language data taken from the discourse in a principled way and compiled into a corpus"(4, 4).

As most of the mentioned linguists asserted, **we also consider** Corpus linguistics a set of methods of analyzing the large collections of texts. By this we mean that this science is *a methodology* which is used for working out the statistics on the target language and for deriving the theory from these data with the help of a bunch of methods. In fact, we are dealing with the Corpus linguistics in order to investigate it application in different aspects of linguistics, and here, we illustrate it as a *methodology*.

Now, let's discuss the other mentioned problems of corpus linguistics. In order to *classify the language corpora into types* the scholars investigated their principles. These include among many others:

- ***properties of the speake****r*: dialect, socio-economic factors, age, gender, education, etc. Labov (1972, 2001)
- ***mode of communication***: spoken, written, computer-mediated, etc. Siekmeyer (2011);
- ***purpose of the communication***: text type, functional variation, register, etc. Biber (1995, 2006).

Now, we think that it would be better to inform about some problematic features of CL. Here, we have analyzed the *disadvantageous* sides of Corpus linguistics. By this we mean that *Corpus linguistics is not able to:*

✓ *provide negative evidence:* this means that a corpus cannot tell us what is possible or correct or not possible or incorrect in language; it can only inform us what is and is not present in the corpus.

✓ *explain why:* CL cannot explain why something is the way it is, only tell us what it is. To find out why, we, as users of language, use our intuition.

✓ *provide all the possible language at the one time:* By the definition, a corpus should be principled: "a large, *principled* collection of naturally occurring

texts…" meaning that the language that goes into a corpus is not random, but planned. However, no matter how planned, principled, or large a corpus is, it cannot be a representative of a language.

It is time to discuss the next challenge for CL. This is the problem of *authenticity* in the language data supplied by corpora. It is often argued that corpora provide learners with 'authentic' or 'real' language, and since these words echo the key features of Communicative Language Teaching (CLT, hereafter) method that favors the use of authentic and real language over concocted ones, it is often assumed that corpus-based language materials are well-suited for CLT. However, some of the researchers have cast doubt on whether language data in corpora are truly authentic. Widdowson contrasted the concept of 'genuineness' and 'authenticity' and argued that 'genuineness' is the property of texts and is an absolute quality, while 'authenticity' is the characteristic of discourse interpretation. He claimed that language in corpora can be genuine, but it is not authentic because it is isolated from discoursal and communicative nature of language.

We consider that it would be relevant to give some information about the following challenging issue. That is how *to measure the proportion* that dialogs make up of the speech of one particular group, for example, adolescents. Corpus compilers can only record a tiny sample of all adolescents, and how would they measure the proportion of dialogs – in terms of time? in terms of sentences? in terms of words? And if they tried to compile a corpus representative of a language as a whole, then how would they measure the importance of a particular linguistic variety? As we can see that a corpus is not always a reliable database of a language or a sublanguage in terms of the mentioned problematic items.

In addition to that, one of the biggest dysfunctions of corpora can be seen from the following quote: "I don't think there can be any corpora, however large, that contain information about all of the areas of [English] lexicon and grammar that I want to explore [...]"(2, 35). It is of course true that the sheer *volume of natural language* will never be able to be *captured* inside a database because it is truly mathematically infinite.

In conclusion, it would be essential to note that every science in its emergence experiences some problems. In fact, Corpus linguistics also has several challenges as mentioned above. However, those have had no proper solutions yet. As new investigators in CL, we believe that there will be undertaken enough researches in order to sort the problems out.

REFERENCES

1. Sinclair J. 1991. Corpus and text-basic principles. – Oxford: Oxford University Press.
2. Fillmore Ch. Corpus linguistics or Computer-aided armchair linguistics// J. Svartvik (ud.) Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, (1992), 35–60.
3. Ганиева И.Ф. Об использовании корпусов в лингвистических исследованиях// Филология и искусствоведение. Т.12. – БГУ, Россия, 2007, № 4. С.104–106.
4. Teubert W. 2005. My version of corpus linguistics. International Journal of Corpus Linguistics 10(1): 1–13.
5. Thompson, Geoffrey and Susan Hunston (eds.). 2006. System and corpus: Exploring connections. – London: Equinox.
6. Tognini-Bonelli, Elena. 2001. Corpus linguistics at work. – Amsterdam: John Benjamins.
7. Stubbs, Michael. 1993. British traditions in text analysis: From Firth to Sinclair. In M. Baker, F. Francis and E. Tognini-Bonelli (eds.). Text and technology: In honour of John Sinclair, 1–36. Amsterdam: John Benjamins.
8. McEnery, Tony, Richard Z. Xiao and Yukio Tono. 2005. Corpus-based language studies: An advanced resource book . London: Routledge.
9. McEnery, Tony and Andrew Wilson. 1996. Corpus linguistics. Edinburgh: Edinburgh University Press.
10. Gries, Stefan Th. 2006b. Introduction to S. Gries and A. Stefanowitsch (eds.).
Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis, 1–18. – Berlin, Heidelberg, New York: Mouton de Gruyter.
11. Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In J. Svartvik (ed.). 105–122.
12. Mahlberg, Michaela. 2006. Lexical cohesion: Corpus linguistic theory and its application in English language teaching. International Journal of Corpus Linguistics (3): 363–383.
13. Chomsky, Noam. 2004. (Interviewed by Andor, Jozsef). The master and his performance: An interview with Noam Chomsky. Intercultural Pragmatics 1(1): 93–111.