International Journal of Multicultural and Multireligious Understanding

http://ijmmu.com
editor@ijmmu.com
ISSN 2364-5369
Volume 8, Issue 4
April, 2021
Pages: 525-531

Corpora and Corpus-Based Teaching Uzbek to Foreigners

Saodat Abdurakhimovna Adilova

Candidate of Pedagogical Sciences, Associate professor, Tashkent State Pedagogical University named after Nizami, Uzbekistan

http://dx.doi.org/10.18415/ijmmu.v8i4.2622

Abstract

The article discusses language corpuses, the history of their creation, requirements for creating a corpus, the possibility of using the corpus in teaching languages. It also examines the current state of the creation of Uzbek language corpus, the problems associated with its creation, and provides examples of using the corpus in teaching Uzbek as a foreign language.

Keywords: *Corpus Linguistics; Corpora; Language Teaching Methods; Corpus-Based Teaching; Authentic Materials; Computerized Language Teaching; Concordance; Language Variability*

Introduction

The process of globalization, which is taking place all over the world, has embraced the educational system of many states. As the need for mutual communication between representatives of different nationalities increases, so does interest in learning different languages.Nowadays, the number of foreigners who want to learn Uzbek language is increasing day by day. First of all, their attention is attracted by the geopolitical status of the Republic of Uzbekistan, historical monuments located on its territory, innovations in the economy, foreign policy, and the social sphere. On the other hand - the customs, traditions of the people living on Uzbekistan, the uniqueness of Uzbek language, the variety of its vocabulary, colorful intonation, the endless possibilities of syntactic synonymy. All this works as a motivation for diligent study of the state language of our country by foreign citizens. In this regard, there is a need to find new effective methods of teaching Uzbek as a foreign language. Certainly, it is difficult to overestimate the role of the comparative typology of languages in this process as well as other main approachs of teaching non-native language. However, in the world there is such a large-scale didactic tool that helps the studying and teaching languages as a linguistic corpus, the creation, development of which is one of the urgent problems of both theoretical and practical linguistics.

In the process of teaching Uzbek as a foreign language, a teacher often faces a number of problems associated with a lack of text material that meets all modern requirements. This is especially true when it comes to an non-turkic-speaking audience. for example, differences in the origin and morphological features of the English and Uzbek languages raise many questions among students, both in terms of semantics and in terms of the logical construction of the sentence, the entire text. In this regard, the

teacher needs a lot of time to prepare new material that helps to maintain interest, increase the motivation of students to learn Uzbek.

Results and Discussion

The corpus of linguistics is one of the sections of linguistics; it is engaged in the collection and processing of corps of texts. This term was introduced into science in the 1960s in connection with the beginning of the creation of corpora, and since the 1980s, thanks to computer technology, this area began to gather momentum [7]. Although the methods used in corpus linguistics were first adopted in the early 1960s, the term "corpus linguistics" didn't appear until the 1980s. Corpus linguistics originally appeared on the basis of the English language, but in a short time ones of other languages began to appear. In 1963, at the University of Brown USA, such scientists as W. Francis and G. Kucher created the first corpus of text, consisting of 500 texts [1]. Each of these texts had two thousand words, the texts were written in the 15 most popular prosaic styles recognized in the USA at that time. In addition, a frequency indicator and some statistics were attached to this corpus.Below are some definitions of the corpus given by various publications and authors.

Doctor of Philology, professor, academician of the Russian Academy of Sciences Vladimir Plungyan explains the essence of this concept as follows: "The corpus of language and the science that is associated with this, corpus linguistics, is such a topic, an area that very quickly burst into the life of linguists in approximately late XX - early XXI century. If we want to name a field of linguistics that is ultramodern by definition, then the first thing that comes to mind is just the linguistics of the corpus"[14:3].

Richard Nordquist who is a freelance writer and former professor of English and Rhetoric who wrote college-level Grammar and Composition textbooks, gives the following description: "Corpus linguistics is the study of language based on large collections of "real life" language use stored in corpora (or corpuses) - computerized databases created for linguistic research. Also known as corpus-based studies [7:1].

In work "Corpus Linguistics and Linguistically Annotated Corpora" the authors Sandra Kuebler and Heike Zinsmeister wrote that corpus linguistics was viewed by some linguists as a research tool or methodology, and by others as a discipline or theory. They conclude "the answer to the question whether corpus linguistics is a theory or a tool is simply that it can be both. It depends on how corpus linguistics is applied" [6:1].

N.V. Kozlova under the corpus means an electronic complex of texts in one or several languages, created on the basis of certain parameters. In her opinion, being a collection of texts in written or verbal form, the corpus differs from ordinary texts in that it is digitized, that is, texts are analyzed electronically, have special tags, they are stored electronically, have linguistic annotations that put all this information in order [13:79].

Hans Lindquist is professor of English Linguistics at Malmo University, Sweden. He has taught English in Switzerland, the United States, Japan, New Zealand and the Gambia and co-edited volumes on translation theory, the major varieties of English, and corpora and grammaticalization. Here is Hans Lindquist's opinion given in his "Corpus linguistics and the description of English" (Edinburgh University Press, 2009): "Corpus linguistics is . . . a methodology, comprising a large number of related methods which can be used by scholars of many different theoretical leanings. On the other hand, it cannot be denied that corpus linguistics is also frequently associated with a certain outlook on language. At the center of this outlook is that the rules of language are usage-based and that changes occur when speakers use language to communicate with each other. The argument is that if you are interested in the

workings of a particular language, like English, it is a good idea to study language in use. One efficient way of doing this is to use corpus methodology . . .." [7:2].

We believe that the definition given by V.P. Zakharov can be considered the most complete:

Linguistic corpus of texts is a large, electronically presented, unified, structured, tagged, philologically competent array of language data designed to solve specific linguistic problems [12:3].

In general, the following qualities are indicated in each description:

1. It is necessary to provide many texts (on the Internet or on disk).

2. The material must be authentic, not edited, it must show the variability of the language units, it is possible to include materials of "live speech".

3. To carry out linguistic analysis, language units must be tagged.

4. As a result of the analysis, it should be possible to distribute language material based on various principles (for example, by genre, date of creation of the text, topic, and so on) [8:171].

So how to use the linguistic corpus in the process of teaching language? Where to begin? Should one study grammar in advance? How rich should a student's vocabulary be in order to be able to understand authentic oral or written texts in the target language? Corpus-based teaching can answer to this kind of questions in new way. The peculiarity of this approach to language learning is that it is proposed to study not the rules of the language, but the texts themselves and examples of the use of language units. And as processing, analyzing texts, a person understands the lexical and grammatical essence of the language.

As Elena Tognini-Bonelli writes: "In the context of the classroom the methodology of corpus linguistics is congenial for students of all levels because it is a 'bottoms-up' study of the language requiring very little learned expertise to start with. Even the students that come to linguistic enquiry without a theoretical apparatus learn very quickly to advance their hypotheses on the basis of their observations rather than received knowledge, and test them against the evidence provided by the corpus" [9:224]. ""To make good use of corpus resources a teacher needs a modest orientation to the routines involved in retrieving information from the corpus, and—most importantly—training and experience in how to evaluate that information" [5:1].

Over the past 10–15 years, researchers in teaching foreign languages have used extensive text corps to assess the realities of the language in its natural state. These corpora of texts have significantly influenced the improvement of the quality level of the published language manuals. Instead of traditional guidelines on how to use the language correctly, new corpus studies describe, empirically reasonably analyze what people really say. New dictionaries created using corpus linguistics techniques such as Longman, Oxford, Collins, as well as the experience of critically rethinking the postulates of descriptive English grammar (Longman Grammar of Spoken and Written English, published in 2000) deserve special mention. The experience gained, as well as generally accepted practice, suggests that the most stable results are obtained by the step-by-step implementation of these new techniques in the learning process to ensure effectiveness and consistent motivation. Thus, at the initial stage, carefully selected quotes should be used in handouts with well-designed instructions and assignments, and at subsequent stages, students will be able to adequately cope with the unpredictability of a "live" search for concordances on the Internet and independently formulate research tasks. The most interesting in this aspect are the critical analytical works on modern English grammar, obtained in the field of natural spoken language, which in

fact reveals much more extensive discrepancies with the standard written language than was ever indicated in the teaching aids. Now, having been identified with the help of case-based techniques, these features can be taken into account in the teaching process and in the development of modern educational and methodological complexes.

In the last decade, another new and extremely promising direction has arisen in the organization of the process of teaching foreign languages, in which the student has the opportunity to resort to using "raw" language data directly from the corps. This area is called data-based learning, or data-driven learning. It is based on solid empirical evidence that students can learn language much more effectively when the use of the observe - hypothesize - experiment model, i.e. when they have the opportunity to make their own conclusions regarding the meanings of words, phrases, grammar rules based on authentic language material. This inductive method complements the more common deductive approach, also known as listen-to-practice-speak, in which students gain knowledge of the rules and definitions from instructor explanations and references. The process is not necessarily limited to a computer terminal. The results of corpus searches (concordances) in printed form can be easily incorporated into handouts, teaching aids, etc. and used in the process of traditional teaching in the lesson.

The purpose of the language corpus is to show the functioning of linguistic units in their natural contextual environment. Based on the corpus, one can obtain data: on the frequency of word forms, lexical units, grammatical categories; about changes in frequencies; about changes in contexts at different time periods; the behavior of linguistic units of different authors; the joint occurrence of lexical units; about the features of their compatibility, management, etc.

It is necessary to consider the issue of text tagging, since the presence of a text bank is not enough to solve language problems. Tagging, annotation is the assignment of special marks to texts and their components. They distinguish between extralinguistic (meta-marking-information about the author and information about the text: author, title, year and place of publication, genre, subject); structural (chapter, paragraph, sentence, word form); linguistic proper. Linguistic markup can be of the following types: part-of-speech tagging, syntactic, semantic, prosodic tagging, etc. [12:3]. By the way, morphological marking is the most widespread. In this case, not only the properties of the parts of speech are taken into account, but also information on various grammatical categories is carefully studied. Morphological tagging is carried out using special automatic programs of morphological analysis. And the syntactic markup provides an explanation of the structure of each sentence using signs, but this requires a lot of time and effort. In addition, as noted earlier, elements of the corpus can have semantic, prosodic, graphic and other tags.

Foreign experience shows that with the creation of electronic corpuses, the possibilities of demonstrating the diversity of linguistic units and the language as a wholesystem, as well as the possibilities of philological and linguocultural studies, integrative language teaching have increased significantly.

Working with corpuses begins with an introduction to the Sketch Engine.SketchEngine is a corpus manager andtext analysis software developed by Lexical Computing Limited since 2003. Its purpose is to enable people studying language behavior (lexicographers, researchers in corpus linguistics,translators or language learners) to search large text collections according to complex and linguistically motivated queries. Sketch Engine gained its name after one of the key features, word sketches: one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behaviour. Currently, it supports and provides corpora in 90+ languages[3].

So how exactly can Uzbek instructor use the corpus while teaching languages? Corpora can help in the following cases:

- in determining the lexical minimum, ie the minimum number of words that students need to know to master the language at a particular level. To understand how necessary is for students to know exactly these words, what is their frequency, which words are outdated and which are considered neologisms;

- in choosing theoretical material for writing textbooks or for a regular lesson. Answer the questions - how important are these grammatical rules, when exactly it is necessary to study which rule, how often there are exceptions to the rules; do the norms of the literary language change, if so, how and for what reason;

- in demonstrating linguistic realities, in order to show how a particular word or phrase actually behaves in life, in which context the studied language units can most often be found, when explaining the valence of words, their compatibility. In demonstrating style differences and syntactic synonymy;

- ihen teaching a language for special purposes, that is, to know what medical students, engineering students, businessmen and others should pay attention to.

- in using an inductive approach to teaching. Students can make certain conclusions by studying examples of concordance. In some cases, this is much more effective than doing exercises to activate vocabulary or grammar rules or reading artificially created texts.

- in analyzing the teaching process and preparing the lesson plans in order to increase the effectiveness of classes.

Here is an example of using the corpus. The one who is studying a foreign language independently reads the text (or piece of text) given in corpus. He formed an assumption about the role that this or that unit of a given text plays, the first guesses and conclusions appeared. By studying the nextpiece of text, the student can come to more confident conclusions and, thus, using the inductive method, the student will gain certain knowledge and skills that will be stored in his memory much longer than the data obtained in traditional textbooks.

Another example: a student needs to understand the difference between almost similar words. He turns to the materials of the corpus, in a second he finds at least 20-30 examples from “live speech”, with the help of contextual analysis he will find out the semantic difference between similar words.We can assume another case. Very often foreign studenthears the word “bir” in various combinations in the speech of native Uzbek speakers. But he cannot determine the specific meaning of this word in a certain context. Many textbooks abound with artificial examples, while grammatical and stylistic descriptions are based more on the intuition of their compilers or on secondary sources. The language corpuswould help a student in a short time to understand that the word “bir” is polysemantic and it expresses different meanings: it can be used both separately and in combination with suffixes,as well as other words, or in pairs, in repeated form. In addition, in oral speech sufficiently large spectrum of expressivity can be obtained by a long pronunciation of a vowel in this word (“b-i-i-i-r”) and changes in intonation [11:125-128]. The student can also be given the task of finding 20 sentences, where the word “bir” expresses the meaning of approximation, another 20 sentences with an expression, emphasis, etc. Thus, the problem associated with the real use of the polysemantic word would be solved.

Modern computerized education is moving away from the just presentation of teaching material. Methods requiring a creative approach and the formation of a culture of self-education are becoming increasingly relevant. As Carol Chapel says, “…in the framework of the course of grammatical analysis, we involve students in the analysis of the computer corpus, which, firstly, affects their perception of the analysis itself, and secondly, their ability to independently conduct such analysis and, thirdly, how they will teach grammar by themselves ”[2:9].

Obviously, the quality of a corpus depends on its creators, on the amount of base material, on the time of its creation. Unfortunately, for many years, the national corpus of the Uzbek language did not exist. Only the corpus created in 2012 was available on the Internet. Below is given information about it:

"The Uzbek Web Corpus (uzWaC) is an Uzbek corpus made up of texts collected from the Internet. The corpus was prepared according to standards described in the document "A Corpus Factory for Many Languages (Kilgarriff et al. at LREC 2010). Data were downloaded in January 2012 with the total size 18 million words. Texts were cleaned and deduplicated. A complete set of tools is available to work with this Uzbek corpus to generate:

- word lists - lists of Uzbek words organized by frequency;

- n-grams - frequency list of multi-word units;

- concordance - examples in context" [10].

Trial use of this case showed that the creation of the case is a painstaking and large-scale work, the quality of which is influenced by each element of the material. The corpus in question is written in the Uzbek Latin alphabet. In all parts of the corpus, except the frequency list, diacritical marks are not recognized by the computer program. Words containing such sign <'> are divided into parts and these parts of words behave as separate tokens. For example, when identifying the compatibility of the word "baland", variants that did not exist in the language were encountered. As examples from some context the corpus showed such word collacations as "yibaland", "li balandkeldi", "baland ko'", and so on. In fact, these should be word combinations "yibaland", "li balandkeldi", "baland ko'". The problem is related to the presence of a special element in the letters o‘ and g‘. Also, the problem will exist if use the Cyrillic alphabet, since in the Uzbek Cyrillic there are also some elements that are not recognized by a computer program. These thoughts, in turn, prove the need to revise and somewhat change the current Uzbek alphabet.

Conclusion

In the world practice of language teaching, corpus methods are recognized as highly effective innovative methods. These methods have aspects of intersubject integration, empirical adequacy, authenticity, flexibility, adaptability, "discovery" in language learning.Over the past years, scientists of Uzbekistan have been actively working on creating the national corpus of Uzbek language. In particular, researches have been preparing materials for new corpora, studying the foreign experience of creating a linguistic corpus, working on solving machine translation problems. Linguists-methodologists of Uzbekistan and foreign scientists have been creating corpuses of individual works of fiction, individual genres, compiling frequency lists. And just a couple of days ago it was announced that a national corpus of Uzbek language which includes more than 100,000 texts has been created [15]. It is also gratifying that a number of microcorps are being formed, such as educational corpus, dialect corpus corporations, author's corpus corporations, poetic corpus corporations, oral, scientific, official corpus corpora, and newspaper corpus corporations. Obviously, this is the result of well-coordinated work of a large team of linguists, experts in literature, IT programmers, lexicographers, methodologists and experts in other spheres who are ready to update, evolve and enrich the corpus constantly. This suggests that teachers, learners and others who are interested in language will be able to use Uzbek corpora very soon. And after all, being familiar with corpuses should be a standard characteristic of both today's language instructors and students.

References

1. Brown Corpus.http://clu.uni.no/icame/brown/bcm.html#bc3
2. Chapelle,Carol A. // Essential teacher. 2003. Vol. 9. P. 5–11.
3. https://en.wikipedia.org/wiki/Sketch_Engine
4. https://ru.m.wikipedia.org.wiki:1
5. John McHardy Sinclair, How to Use Corpora in Language Teaching, John Benjamins, 2004. https://www.thoughtco.com/what-is-corpus-linguistics-1689936
6. Kuebler S., Zinsmeister H., Corpus Linguistics and Linguistically Annotated Corpora, 2015 https://www.bloomsbury.com/us/corpus-linguistics-and-linguistically-annotated-corpora-9781441164476/
7. Nordquist R., Definition and examples of corpus linguistics. https://www.thoughtco.com/what-is-corpus-linguistics-1689936by Updated July 03, 2019.Nordquist R., Definition and examples of corpus linguistics. https://www.thoughtco.com/what-is-corpus-linguistics-1689936by Updated July 03, 2019.
8. Sinclair J. Corpus, Concordance, Collocation. Oxford, 1991. –P. 171.
9. Tognini-Bonelli E., Corpus Linguistics at Work. John Benjamins, 2001. xii, 224 pp. https://benjamins.com/catalog/scl.6
10. uzWaC: Uzbek corpus from the webhttps://www.sketchengine.eu/uzwac-uzbek-corpus/
11. Жамолхонов, 2004:125-28 // Жамолхонов Ҳ. Ҳозирги ўзбек адабий тили. –Т.: ТДПУ, 2004.
12. Захаров В.П. Корпусная лингвистика. СПб., 2005 –С. 3////Захаров В.П. Корпусная лингвистика. СПб., 2005 –С. 3
13. Козлова Н.В.Лингвистические корпуса: определение основных понятий и типология. Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2013. Т.11 Выпуск 1. – С. 79.
14. Плунгян В. Что такое корпус языка? https://postnauka.ru/video/7783
15. http://til.gov.uz/uz/news-and-announcements/news/237