CREATION OF THE EDUCATIONAL CORPUS OF STUDENT SPEECH IS AN EXAMPLE OF WORLD EXPERIENCE

Kuvondikova Gavkhar Isomiddinovna
Teacher of the National University of Uzbekistan.

Urinboeva Nazokat
Master's degree student
of the National University of Uzbekistan.

Abstract: This article presents the importance of the educational corpora of students' speech and the analysis of corpora samples created in this regard. Extensive work has been done on Arab, Brazilian and Lithuanian student corpora. Examples of them include the development of criteria for teaching the language as a mother tongue, as a second and foreign language, and the creation of statistics on the use of academic words by students. Adaptation of texts or spoken speech samples (audio or multimedia files) attached to the corpus content for the purpose of the user, as well as the important aspects of following the rules of providing metadata in order to make the corpus easier to use are expressed in the article.

Keywords: Educational corpus, Arab student corpus, Brazilian student corpus, Lithuanian corpus

Аннотация: В данной статье представлена значимость учебного корпуса речи студентов и проведен анализ образцов корпусов, созданных, в связи с этим. Обширная работа была проделана над студенческими корпусами арабских, бразильских и литовских стран. К их примерам можно отнести разработку критериев обучения языку как родному, как второму и иностранному, а также создание статистики использования академических слов учащимися. Адаптация текстов или образцов устной речи (аудио или мультимедийных файлов), прикрепленных к контенту корпуса, для целей пользователя, а также важные аспекты соблюдения правил предоставления метаданных с целью облегчения использования корпуса выражены в статья.

Ключевые слова: образовательный корпус, арабский студенческий корпус, бразильский студенческий корпус, литовский корпус.

The growing public interest in the use of corpora as a resource in the study of language and linguistics led to the creation of TaLC in 1994 as a result of a discussion among members of ICAME (the international computer archive of modern and medieval English). . The first conference on educational corpora was held in Lancaster in 1994, and it was noted that the use of computer text corpora in research was well established and that they were widely used for teaching purposes, corpora data in the preparation of teaching materials, and foreign languages and discussions on the direct creation of corpora by students in linguistic education are on the agenda.

Creation of the first educational corpora Sylviane Granger and her scientific team created the "International corpus of learner English" (ICLE - International corpus of learner English). The purpose of this corpus is to create methods for correcting common errors of language learners by classifying them [D. Stewart: 2004].

The following characteristics of English language learners were taken into account when creating this educational corpus [G. Sylviane; 199]: educational environment, age, mother tongue, stage of education, nature of tasks.

Language learners were selected for this corpus within the following criteria:

1) type of learners;

2) age: adults

3) level of education: high level (advanced);

4) task: essays

The educational international corpus of the English language is based on the students of English as a second language of French, German, Chinese, Japanese, and Czech nationalities.

The international educational corpus of the English language consists of several linguistic instruments.

• The ICE Markup Assistant is an automatic system that simplifies the process based on a set of standard symbols for this instrumental layout;

• TAGSELECT – the Ice tag Selection system (tagging system) is an automatic sorting system of word group tags generated using the TOSCA tagger for each word in the corpus;

• ICEMARK – Used for syntactically parsing the tagged text before the TOSCA parser. Also, one of its other functions is the ability to change the tags of word groups, edit the errors of the words found in the text, and see untagged sentences.

• ICECUP – Corpus Utility is functionally important for analysis, concordance and search of this instrumental corpus.

Student corpus research is a relatively young but high-performing field of corpus linguistics, and the world experience is rich in extensive and diverse research in this field. One of these is the ALC project, a collection of written and spoken materials collected from Arabic language learners in Saudi Arabia. The aim of this project is to serve as a source of information for the study and teaching of Arabic as a second language and for further linguistic research.

Prior to the creation of the ALC corpus, several studies on student corpora were conducted in Saudi Arabia. These are: Pilot Arabic Learner Corpus (Abuhakema et al., 2008), Corpus of L2 (Hassan and Daud, 2011). The Pilot Arabic Learner Corpus (PALC) is a corpus of approximately 9,000 words of Arabic texts created by American English speakers learning Arabic. Some of the student texts were written while the students were studying Arabic in the United States. This is considered to be an intermediate level and consists of 3818 words. The second part, about those who went to study in Arab countries, was written in Arabic, and it was a high level and made up 4741 words. Corpus of L2 is a corpus of Malaysian Arabic language learners and consists of students' written texts. This corpus is mainly designed to check the incorrect use of Arabic compounds. It contains approximately 240,000 words created by students from 60 universities. 97% of them are Malaysians and were collected in the first and second year of Arabic studies. Corpus materials were developed on the topic of descriptive and comparative essays.

Based on data from both corpora, a further project ALC was created about the student corpora [Abdurakhmonova, N. 2019,2021,2022]. This corpus was more detailed and comprehensive than the corpus created before. First of all, special attention was paid to the body design. In this case, the ALC is intended to consist of a number of small cases. The design of the body is divided according to the following criteria:

non-native speakers of Arabic	students whose native language is Arabic
Men	Women
Pre-university students	university students
oral materials	written materials

The information in the ALC corpus is both oral and written, both of which have two stages. For the first stage of the oral part, the conditions of "describing a photograph" and for the second part of the "interview" have been set. For the first stage of the written part, the topic "Holiday Adventures" was chosen. Students wrote an essay in the classroom for 40 minutes. At the next stage, the students will discuss the content of the discussion "Should students choose education based on their interest?" Or the bright developing areas of the economy?" wrote an essay on the topic. Only the students had to write the essay for two days as homework, not in the classroom or in the auditorium. During the essay, students are given the opportunity to use dictionaries and grammatical information. ALC metadata is classified as:

Reader metadata Text metadata

1. Age 1. Place of registration
2. Gender 2. Year of writing
3. Nationality 3. City of registration
4. Mother tongue 4. The time is fixed
5. The number of languages you can speak 5. The time is not set
6. Number of years studying Arabic 6. Use of literature
7. Number of years spent in Arab countries 7. Use of grammar books
8. General level of education 8. Use of monolingual dictionary
9. Level of education 9. Use of other resources
10. Year/Semester 10. Using a bilingual dictionary
11. Educational institution 12. Text volume

In Saudi Arabia, except for pre-school educational institutions, all other educational institutions are taught in single-sex classes, that is, men and women do not mix. For this reason, special attention is paid to the gender component in the research. To work with women, only female specialists were included. Therefore, the part devoted to information about women's speech was 20%. Especially the oral part of the corpus is laborious and time-consuming work, which made up only 10% of the ALC.

The distribution of words in 4 groups that make up the ALC criterion was as follows.

The level of the students' mother tongue	**The level of education of the students**	**The gender of the students**	**The form of the material**
National 50% 100,000 words	Pre-university students 70% 140,000 words	Male 80% 160,000 words	Written speech 90% 180,000 words
Non-national 50% 100,000 words	University students 30% 60,000 words	A woman 20% 40,000 words	Oral speech 10% 20,000 words

The ALC corpus includes 1,585 texts (written and spoken), 282,732 words developed by 942 students. The corpus differs from other corpora in that it is based on metadata. According to

the metadata, the age of the students is from 16 to 42, and the students of 67 different nationalities participated. Text metadata is classified as follows. In ALC, two genres constituted narrative (67% of ALC content) and discussion (%). According to where the texts were written: students wrote their texts in class (62% of ALC data) at home (31%), but all audio recordings were produced in class (7%). The average length of texts is 178 words33

Another major study of corpora of student speech was conducted in Brazil. Researcher Larisa Goulart Da Silva, a professor at the University of Warsaw, focused her research on how Brazilian students use academic words and how corpus linguistic tools work based on AWL, a pre-existing list of academic words. Also, this study can be a source for creating a list of academic vocabulary used by Brazilian students. For this, the criteria for receiving essays from Brazilian students studying in Britain has been developed [Abdurakhmonova, N. 2019,2021,2022]. Essays were collected in two stages, the essay topic in the first stage was descriptive, and in the second stage, essays were discussed. Larisa Goulart Da Silva initially tried to find answers to the following questions in the students' written works.

a) What is the vocabulary of essays written by Brazilian students?

b) How does it compare with the vocabulary profile of other academic British corpora?

c) What words do Brazilian students use in the Academic Word List (AWL)?

d) What are the differences between Brazilian students' use of academic words and those represented in the British Academic Written English (BAWE) corpus?

The texts collected in the corpus are classified based on the following groups:

- Life Sciences,
- Social studies,
- Arts and humanities,
- Physical sciences.

In the section of groups, we can see the level of students' use of academic words in the table below.

T|r	Corpus	Number of running words	AWL
1	Art	875.000	9.30%
2	Science	875.000	9.10%
3	Commerce	875.000	12%
4	Law	875.000	9.40%
5	Linguistics	2.031	12.60%
6	Applied linguistics	5.137	17.60%
7	Sociology	2.084	12.44%
8	Social psychology	2.059	14.38%
9	History	2.036	14.49%
10	Devolepment	2.023	12.26%
11	Medicine	2.024	6.72%
12	Zoology	2.026	7.31%
13	Anatomy	11.356	8.60%

The researcher worked in cooperation with the "Languages Without Borders" (SwB) program. The program covers 10,740 Brazilian students from 87 British universities. These students came from 131 Brazilian universities representing different regions of Brazil. According to information provided on the official SwB website, of these 10,740 students, 774 are doctoral

students, 571 are full-time PhD students, 8,864 are visiting undergraduate students, and 531 are postdoctoral fellows. The researcher recruited only undergraduate students for the corpus. The results of the study show that Brazilian students use academic vocabulary at the same level as British students, but there are some peculiarities in the use of academic vocabulary by Brazilian students. Given these unique characteristics, some pedagogical recommendations for teaching EAP to Brazilian students are presented below. Nevertheless, the recommendations given here mainly take into account the "Languages Without Borders" program. For example, BAWE students used 43 prefixed word forms, while Brazilian students used only one prefixed word form, which means that Brazilian students used less suffixes and prefixes than BAWE students. Because of this, it has been hypothesized that prefixes such as "under" and "over" may not be comfortable for Brazilian students' speech because they are not derived from Latin. However, further research is needed to confirm this hypothesis.

Student corpora are important in students' second language acquisition. In Lithuania, a corpus of EFL students was created in 2011. Students of the Faculty of Philology of the English-speaking Vilnius University in Lithuania participated in the study[Abdurakhmonova, N. 2019,2021,2022].The corpus focused on examining the phraseology of student speech and various aspects of academic written and spoken English. Materials in EFL are classified under the following groups:

Local - Non-local

Women – Men

The oral part of the research begins with an informal discussion about university life, hobbies, travel or the future. Then interviewees are asked to choose one of three topics: "The country that surprised you", "Your favorite movie to watch", "The game you liked or disliked". Each interview ends with a short story based on a picture. Interviews averaged 200 words. In the written part, students wrote an argumentative essay for 40 minutes. Research shows that there are common mistakes for different groups of students. For example, excessive use of connectives and incorrect use of lexical expressions in the written text, some deviations in the use of grammatical patterns. It has been proven that the reason for uncertainty and hesitation in students' oral speech or time-consuming attempts is that the speech is subject to grammatical rules, or that the students could not speak without a sequence of grammatical formulas.

The educational corpus of the Uzbek language was created within the framework of the practical project "Creation of the educational corpus of the Uzbek language" numbered AM-FZ-201908172, its initial presentation was held on April 23, 2021, and its base is currently being enriched and improved. Today in this case

1) morphoanalyzer (automatic morphological analysis);

2) dividing the word (form) into syllables;

3) provide comment(s);

4) show antonyms;

5) synonymizer (a program to present synonyms to the word entered in the search);

6) presentation of homonyms;

7) giving the pronunciation (paronym) of the word with its explanation;

8) providing information from the "Native Language Encyclopaedia" related to the searched word;

9) display phrases in which the searched word is included;

10) it has the ability to display a range of ratings for various features.

At the moment, the volume of linguistic data of the educational corpus of the Uzbek language has the following indicator:

№	The name of the base unit	Count
1	Books	184
2	Internet texts	4 78 908
3	Phrases are tagged words	36 897
4	Words	5600
5	Contexts	5562
6	Number of words annotated	4575
7	Phrases	1384
8	Word order	413
9	Synonymous words	993
10	Antonym words	870
12	Homonyms	1636
13	Concepts in "Encyclopedic Dictionary of Mother Tongue"	1636

Various studies are being carried out to enrich the educational corpus, to enrich its base with more useful and necessary information for the student. The most popular of these studies is the study of students' speech and corpora. Student corpora research is a relatively young but highly developed field of corpus linguistics.

In conclusion, it can be said that using the positive results and experiences achieved in the educational corpora of student speech studied above, the creation of corpora of student speech in the educational corpus of the Uzbek language and the formation of methodical developments on bub are among the priority tasks in the future.

Foydalanilgan adabiyotlar:

1. Abdurakhmonova N, Urdishev K. Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL). 2019;6(1-2019):131-7.
2. Abdurakhmonova N. Dependency parsing based on Uzbek Corpus. InProceedings of the International Conference on Language Technologies for All (LT4All) 2019.
3. Abdullah Alfaifi, Eric Atw, Hedeya Ibrahim. Arabic Learner Corpus (ALC) v2. A New Written and Spoken Corpus of Arabic Learner2014 Saudia Arabia
4. D. Stewart, S. Bernardini, G. Aston Introduction: Ten years of TaLC / Corpora and language learners, Vol 17, John Benjamins Publishing Company Amsterdam/Philadelphia, 2004, - P. 3
5. Larisa Goulart Da Silva Academic vocabulary: a corpus linguistics study on how Brazilian students write academic English. 2016-a. Brazilia
6. Abdurakhmonova, N., & Urdishev, K. (2019). Corpus based teaching Uzbek as a foreign language. *Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL)*, *6*(1-2019), 131-7.
7. Абдурахмонова, Н., & Абдувахобов, Г. (2021). O ‘quv lug ‘atini tuzishning nazariy metodologik asoslari. *Международный журнал искусство слова*, *4*(6).

8. Abdurakhmonova, N. (2021). Formal-Functional Models of The Uzbek Electron Corpus. *ANGLISTICUM. Journal of the Association-Institute for English Language and American Studies*, *10*(8), 59-66.
9. Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. In *2022 7th International Conference on Computer Science and Engineering (UBMK)* (pp. 73-75). IEEE.
10. Abdurakhmonova, N., Shakirovich, I. A., & O‘G‘Li, K. N. S. (2022). Morphological analyzer (morfoAnalyse) Python package for Turkic language. *Science and Education*, *3*(9), 146-156.
11. Granger, S. (2003a). Error-tagged learner corpora and CALL: A promising synergy. CALICO Journal, 20(3): 465-480.
12. Granger, S. (2003b). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. TESOL Quarterly, 37(3): 538-546.
13. Hammarberg, B. (2010). Introduction to the ASU corpus, a longitudinal oral and written text corpus of adult learners’ Swedish with a corresponding part from native Swedes. Stockholm University:Department of Linguistics.
14. Thompson, P. (2005). Spoken language corpora. In M. Wynne (Ed.), Developing Linguistic Corpora: a Guide to Good Practice (pp. 59-70). Oxford: Oxbow Books.