UZBEK TEXT REPRESENTATION

Mauluda Urazalieva
National University of Uzbekistan
Tashkent, Uzbekistan
urazaliyeva_m@nuu.uz

Shukhrat Kurbonaliev
"Venkon Group"
Tashkent, Uzbektistan
shukhrat0594@gmail.com

Abstract – In the era of modern technology, wide opportunities are being created within the framework of applied linguistics and computer linguistics. In particular, language corpora actively use the capabilities of the audio corpus, which are necessary for speech recognition and improve the quality of voice translation of texts. The article compares the interface, size and focus of the corpora created within the world and turkish languages and also talks about the representativeness of the text, which is important in the audio corpus software that is planned to be created in the uzbek language corpus.

Key words: audio corpus, text representativeness, text sorting, corpus chronology, corpus annotation.

ПРЕДСТАВЛЕНИЕ УЗБЕКСКОГО ТЕКСТА

Мавлуда Янгибаевна Уразалиева
Национальный университет Узбекистана
Ташкент, Узбекистан
urazaliyeva_m@nuu.uz

Шухрат Хушбахтович Курбоналиев
«Venkon Group»
Ташкент, Узбекистан
shukhrat0594@gmail.com

Аннотация – В эпоху современных технологий создаются широкие возможности в рамках прикладной и компьютерной лингвистики. В частности, языковые корпуса активно используют возможности аудиокорпуса, необходимые для распознавания речи и повышения качества голосового перевода текстов. В статье сравниваются интерфейс, размер и направленность тюркских корпусов с

другими языками, а также говорится о репрезентативности текста, что важно при создании программного обеспечения аудиокорпуса, которое планируется создать в рамках корпуса узбекского языка.

Ключевые слова: аудиокорпус, репрезентативность текста, сортировка текста, хронология корпуса, аннотация корпуса.

I. Introduction

World in computer linguistics and corpus linguistics problems education began in the 40s of the XX century. In particular, in the 60s of the 20th century. process accelerates, 21st century in their heads own as part of millions words reflection brought from the face more than language body appear It was. Artificial intelligence auto translation, computer analysis and to its editing, thesaurus, electronic dictionaries like possibilities expanded his scientific-theoretical basics was created and on practice apply possible first samples apply began.

This article one how many or same one in the language texts eventually based electron as assembled data system, audio case in building in the case lyrics too representativeness will try to be and his linguistic supply volume check taken.

II. Literary analysis and methods

Corpus linguistics world computer linguistics majority good developing from the fields one is significant to results achieved Russian linguist from scientists one R.G.Piotrovsky stated : “Probably linguistic data big massive texts only from the complex taken possible” [1]. Next years during research in my work seekers directly body with their work take are going and this linguistics on the field language technology is also involved is being done.

Practical from point of view corpus linguistics in the 60s of the 20th century Brown’s Corpus founders to the basis Brown Corpus is an electronic collection of samples of American-English text, the first large structured corpus of various genres [2]. It was this corpus that began the tradition of free use of the corpus for research. based on this corpus, the American Heritage Dictionary was founded in 1969[3].

The field of corpus research in Turkish studies is developing in scientific research on turkish languages. M. Aksan, D. Zeyrek, K. Oflazar, U. Ozge [4] about *the Turkish language corpus;* L.A.Buskunbaeva,

Z.Sirazitdinov[5] in *the Bashkir language; A. Sheimovich about the Khakass language*; in *the Tatar language* J. Suleymanov, A. Gatiatullin, O. Nevzorova, R. Gilmulin, B. Khakimov [6]; Research was carried out by many scientists, for example L. Kubedinova on *the Crimean Tatar language* and A. Salchak[7] on *the Tuvan language.*

The article analyzes the similarities and differences between the corpora of English, Russian and Turkish languages with the audio corpus that is supposed to be created in the uzbek language. The capabilities of the uzbekcorpus.uz platform are listed.

III. Results and discussion

The British National Corpus (BNC)[8] is one of the largest sample corpora, containing approximately 100 million words. The corpus was created between 1991 and 1994 at the University of Oxford with support from Lancaster University and the British Library. Contains texts of unlimited subject matter and style. All texts in the national corpus were segmented, and the words in the sentences were subjected to grammatical analysis. In addition, the included texts were edited according to three main criteria: the time the text was written, the region and its publisher.

Another well-known corps is the Czech National Corps (Český národní korpus) [9]. This is a modern Czech language, a synchronous morphological, structured corpus. The CHNK Institute was founded in 1994 with the help of the Czech Ministry of Education through the announcement of various grants and sponsorships from the University of Prague. The first sample of this corpus included written texts containing about 100 million words, not counting small colloquial and dialect words.

In Russian majority first corpus in the 1980s, Russian language frequency dictionary based at Uppsala University in Sweden created in this case to the created to Russia in the 1960s and 1970s language 1 million per body. around words basic source by doing received “Russian language” frequency dictionary was created was (L.N.Zasorina, 1977) [1]. The corpus includes various socio-political texts, fiction, scientific and popular science texts. L.N. Zasorina supervised the collection of lexical material for the corpus and the development of the main program.

The Turkish Web Corpus (trTenTen) is a Turkish corpus of texts collected from the Internet. The corpus belongs to the TenTen corpus

family, which is a collection of web corpora with a target size of 10 billion words, all executed using the same method. Sketch Engine currently provides access to the TenTen corpus in over 30 languages. The data was analyzed by SpiderLing between December 2011 and January 2012 and consists of 3.3 billion words across 12 million documents. This Turkish corpus was processed using the TRMorph tool, which provides part-of-speech tags and roots (word parts without affixes).

The Uzbek language electron corpus project, which is in effect today and is expanding every day, was created within the framework of the international project ERASMUS for the period 2018-2021, and the founder of this project is Professor N. Abdurakhmonova, and the corpus is increasing every year. When starting the case in the interface, you can select the language in which the program will work, and on the left side, through the items, select the desired area. When searching for words, you can choose which field to search in, as well as the style, period. For instance *"ona"* according to the lemma and chose the artistic style and the Soviet era. As a result, 137 words were found in 28 documents. This corpus is also searchable using a morphoanalyzer. In this case, each word is subjected to morphological analysis, it is shown which category it belongs to, and is divided into stems and suffixes.

The term "corpus" usually texts collection means from the corpus time pass with volume and compound change but it's possible changes his into the structure effect not to do needed, like this perfect view housing representativeness. Accordingly, any of steel representativeness constant from problems one existence remains being created corpus for linguistic supply in building volume, content, texts sorting and chronology check taken necessary. Volume the problem is common in the 60s and 70s of the 20th century dictionary create in processes evident surface released original size 1 million word units comprises, Brown Hull, Lancaster - Oslo-Bergen Hull And L.N.Zasorina Russian language under guidance frequency dictionary when volume problem collided [1]. Such cases words another including forms need check received and later volume 100 million word around be possible included. But stick out your tongue education for this less was for today come corpus volume billions achieve maybe this one by language another in aspects learn service does

The corpus is not just text. Its multimedia views also contain audio and video versions of texts. Voice programs have long been a part of

everyday life. For example, the job of many smart apps (such as Apple's Siri, Amazon's Alexa, or Google's voice assistant) is to detect the user's surroundings. The keyword recognition system is designed to continuously listen for command tone words when controlling actions such as opening or closing a door or launching another more complex command interface. Although such voice programs are considered a convenience in the information age by some people, they are a necessary aid for people with disabilities. In particular, it serves as an important educational tool for the blind.

IV. Conclusions

The corpus is not just text. Its multimedia views also contain audio and video versions of texts. Voice programs have long been a part of everyday life. For example, the job of many smart apps (such as Apple's Siri, Amazon's Alexa or Google's voice assistant) is to detect the user's surroundings. The keyword recognition system is designed to continuously listen for command tone words when controlling actions such as opening or closing a door or launching another more complex command interface. Although such voice programs are considered a convenience in the information age by some people, they are a necessary aid for people with disabilities. In particular, it serves as an important educational tool for the blind. In order to create an electron corpus of the Uzbek language, first of all, it will be necessary to compile a large amount of texts on various topics. For example, texts about the styles available in the language are selected. Moreover, it is relatively easy to create a literary corpus of fiction texts or author corpora. Because the life of an author and the preservation of his works require great responsibility. Similarity can be found in many corpora, and naturally creating such corpora is useful for tracking the evolution of a language. It is also possible to keep statistics of grammatical, lexical, and morphological words entering or leaving the language. Efficiency increases if the employee of different fields can use such corpuses and can use them in the work process. Creating an electronic database and corpus for languages that are at risk of falling out of use will allow to preserve scientific and literary literature related to this language for years, and to carry out a number of scientific works on them.

References

[1] V.P. Zakharov. Corpus linguistics: A textbook for students of the "Linguistics" direction. 2nd ed., revised. and additional., – St. Petersburg: St. Petersburg State University. RIO. Faculty of Philology, 2013.

[2] https :// variang. Helsinki. fi / CORD / corporation / BROWN /

[3] https://www.ahdictionary.com/

[4] K. Oflazer. Two-level description of Turkish morphology. Literary and Linguistic Informatics, - Vol. 9, No. 2, - 1994.

[5] Sirazitdinov Z.A., Sirazitdinov B.Z. Body project in Bashkir language linguistics. / International Conference Turklang 2013 - P.59 International Conference Turklang 2013.

[6] Suleymanov D., Gilmullin R., Gatauillin A. System of morphological analysis of the Tatar language based on a two-level morphological model / Turklang 2017. Kazan, 2017.

[7] https://factored.ai/2021/12/14/multilingual-spoken-words-corpus-50-languages-and-over-23-million-audio-keyword-examples/

[8] Abdurakhmonova N.Z., Urazalieva M.Yu. Uzbek language electron in the building (http://uzbekcorpus.uz/) orally texts body creations theoretical And practical Problems. Academic research in educational sciences. 2022 http://www.ares.uz/uz/maqola-sahifasi/uzbek-tili-elektron-korpusida-httpuzbekcorpusuz-oral-texts-corpus-of-creating-theoretical-and-practical-issues.

[9] http://modmorph.turklang.neg/uz/statistics

[10] Abdurakhmanova N.Z. Car translation linguistic supply. - T., 2018.

[11]https://uzbekcorpus.uz/rusVer

[12] 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.

[13] Abdurakhmonova, N., Tuliyev, U., Ismailov, A., & Abduvahobo, G. (2022). UZBEK ELECTRONIC CORPUS AS A TOOL FOR LINGUISTIC ANALYSIS. In *Компьютерная обработка тюркских языков. TURKLANG 2022* (pp. 231-240).

[14] Abdurakhmonova, N., Tuliyev, U., & Gatiatullin, A. (2021, November). Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus. uz. In *2021 International Conference on Information Science and Communications Technologies (ICISCT)* (pp. 1-4). IEEE.

[15] Isroilov, J., & Abdurakhmonova, N. (2018). Personal names spell-checking–a study related to Uzbek. *Journal of Social Sciences and Humanities Research*, *6*(02), 1-6.