CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES 2(12): 161-166, December 2021
DOI: https://doi.org/10.37547/philological-crjps-02-12-31
ISSN 2767-3758

Crossref doi Google Scholar
Accepted21th December, 2021 & Published 26th December, 2021

THE IMPORTANCE OF LANGUAGE CORPUS IN THE CONSTRUCTION OF LEXICOGRAPHIC SOURCES

Manzura Abdurashetovna Abjalova

Doctor Of Filology (Phd), Alisher Navo'i Tashkent State University Of Uzbek Language And Literature, Uzbekistan

ABSTRACT

Language corpus is an important tool for language education - the creation of lingua-didactic and dictionaries, as well as for various researches, diachronous and synchronous study of language, the development of speech competence, vocabulary, speech patterns. In particular, the creation of frequency dictionaries based on it allows creating a list of real active words for language learning and translation dictionaries. It also assists to save time and avoid large-scale manual work.
This article discusses the creation of frequency dictionaries on the basis of language corpus and the possibilities of the Uzbek language corpus.

KEYWORDS: - Corpus, lexicography, frequency dictionary, terminological dictionary, educational corpus.

INTRODUCTION

Article was implemented within the framework of the practical project "Creation educational corpus of the Uzbek language" № AM-FZ-201908172

Modern information technology has opened the door to endless possibilities in the use of the functional capabilities of a language. Computer translation, automatic editing and analysis, speech synthesizers that transcribe written text, speech recognition programs that translate spoken speech into written text, electronic dictionaries, mobile linguistic applications, thesaurus (language treasures) and language ontology are proof of our point. In particular, the creation of a culture of modern electronic dictionaries and their use has proven to be effective in acquiring language skills. The role of language corpus, which need to be created on a global scale, is invaluable in language teaching and the creation of electronic dictionaries.
It is a well known fact, that the use of computer technology in corpus linguistics determines the period and frequency of word use on the basis of language corpus, development of terminology, study and analysis of sentence structure, creation of n-language database for translation programs,

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES 2(12): 161-166, December 2021
DOI: https://doi.org/10.37547/philological-crjps-02-12-31
ISSN 2767-3758

Crossref doi Google Scholar
Accepted21th December, 2021 & Published 26th December, 2021

study of methodology, development of electronic linguodidactics. The ability to create dictionaries was transferred from the card index to the automated process, creating unprecedented conveniences [1]. When it comes to compiling dictionaries, the usefulness of language corpus, especially the creation of frequency dictionaries, is important in the formation of terminology in a particular field. Translator dictionaries using the translation database of bilingual or multilingual parallel corpus also serve as a major resource in the formation of linguistic databases of automatic translation programs.

Compiling dictionaries.The corpus is a large source for compiling dictionaries. Over time, the corpus has become a powerful information resource. Computer-generated dictionaries are being created and processed faster than ever before on the base of the corpus. For example, there are several dictionaries based on the Russian National Corpus [2, 3, 4, 5, 6]. It is noteworthy that the glossema (keyword) of electronic dictionaries based on dynamically changing language corpora and the dictionary article races against the time and remains authentic.

Commenting on the popularity of the text corpus, Charles Fillmore said: “I can make two comments. First, I don't think there can be a body of text that contains information on all areas of vocabulary and grammar, no matter how big the corpus is. The second explanation is that every corpus I have read, no matter how small it is, has shown me facts that I would never have been able to find in any other way. ”[7] According to Fillmore, there is no limit to the size and scope of language corpora, and with such unlimited possibilities, the corpus cannot compete with other resources, systems, and software. Therefore, the creation of corpus-based dictionaries reflects the viability of the language and the real vocabulary in the dictionaries.

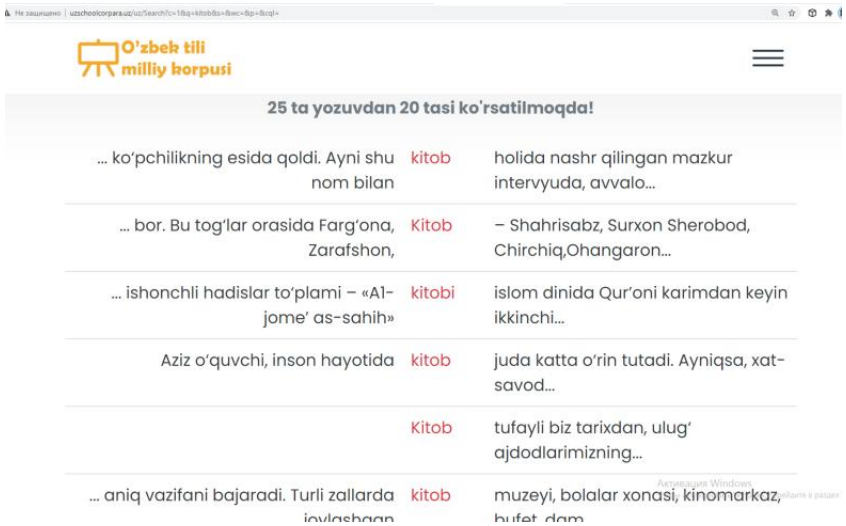
To develop the field of terminology, the terminology of a particular field is to be formed from the texts of millions of different fields in the database. For example, the English-Latin term for engineering, banking, finance, construction, information technology, and international relations can be identified by searching for a single corpus. One of the consequences of the penetration of computer technology into all fields over the last decade is, firstly, the need to create different types of electronic lexicographic sources, and secondly, the electronization of texts in different fields.
There are two reasons for creating corpus-based terminological dictionaries:
1) The corpus contains a large amount of annotated written and oral texts in various fields. This is especially true of the general corpus, the national corpus of natural language. Annotation, that is, the grammatical and semantic interpretation of each word, reveals the interpretation of a particular word in the field, and gives it a general or terminological character.
2) It is possible to distinguish between active and inactive terms by determining the frequency of words used in texts belonging to a particular field.
The set of terminological bases is the basis of any research and is an important factor in the development of computer lexicography.

Structure and prospects of the Uzbek linguistics corpus for education. The case consists of a case interface (1), a search engine (2) and a dictionary section (3). In the "Dictionaries" section there is an annotated dictionary, dictionaries of homonyms, synonyms, paronyms and antonyms as lexicographic resources. The first page of the website of the Uzbek linguistics corpus for education contains basic information about the corpus and its creators, you can go to any page in the menu on the right.
Through search bar, the user identifies:
1) all forms of a particular word with an array of

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES 2(12): 161-166, December 2021
DOI: https://doi.org/10.37547/philological-crjps-02-12-31
ISSN 2767-3758

Accepted21thDecember, 2021 & Published 26thDecember, 2021

examples



Picture 1. An array of examples involving the word "book."

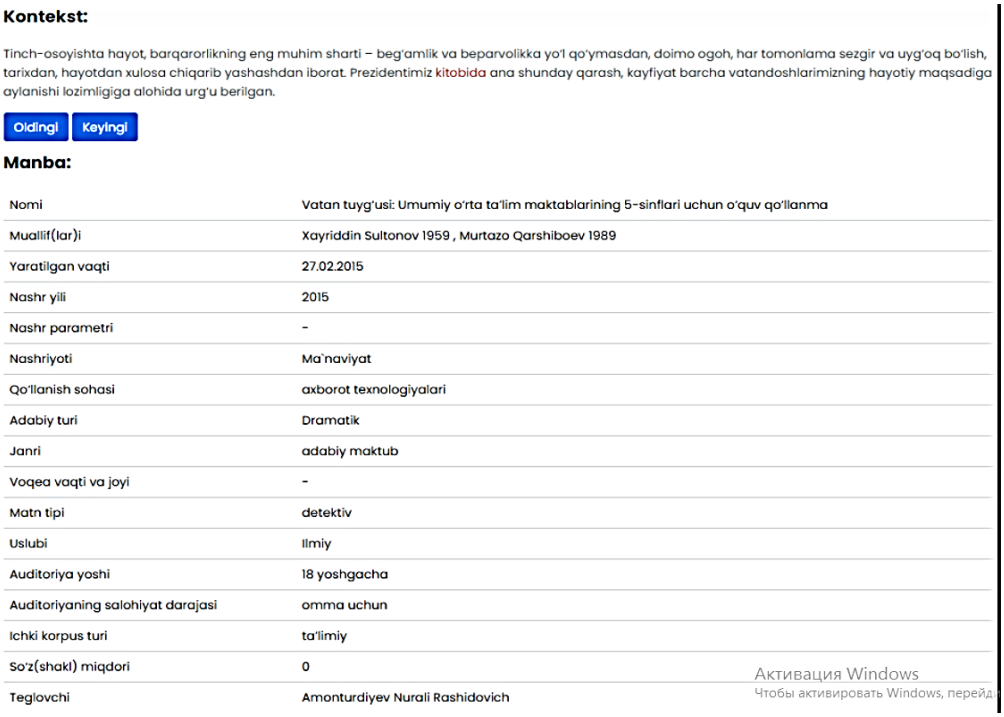
2) identify the source of the examples;

Picture 2. The paragraph in which the word "life" is presented and the name of its source

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES 2(12): 161-166, December 2021
DOI: https://doi.org/10.37547/philological-crjps-02-12-31
ISSN 2767-3758

Crossref doi Google Scholar
Accepted21th December, 2021 & Published 26th December, 2021

3) obtain information about the source; For example:



Kontekst:

Tinch-osoyishta hayot, barqarorlikning eng muhim sharti – beg'amlik va beparvolikka yo'l qo'ymasdan, doimo ogoh, har tomonlama sezgir va uyg'oq bo'lish, tarixdan, hayotdan xulosa chiqarib yashashdan iborat. Prezidentimiz kitobida ana shunday qarash, kayfiyat barcha vatandoshlarimizning hayotiy maqsadiga aylanishi lozimligiga alohida urg'u berilgan.

Oldingi | Keyingi

Manba:

Nomi	Vatan tuyg'usi: Umumiy o'rta ta'lim maktablarining 5-sinflari uchun o'quv qo'llanma
Muallif(lar)i	Xayriddin Sultonov 1959 , Murtazo Qarshiboev 1989
Yaratilgan vaqti	27.02.2015
Nashr yili	2015
Nashr parametri	-
Nashriyoti	Ma`naviyat
Qo'llanish sohasi	axborot texnologiyalari
Adabiy turi	Dramatik
Janri	adabiy maktub
Voqea vaqti va joyi	-
Matn tipi	detektiv
Uslubi	Ilmiy
Auditoriya yoshi	18 yoshgacha
Auditoriyaning salohiyat darajasi	omma uchun
Ichki korpus turi	ta'limiy
So'z(shakl) miqdori	0
Teglovchi	Amonturdiyev Nurali Rashidovich

Picture 3. Metalinguistic information on the literature of the 5th grade "Vatan tuyg'usi"

4) get acquainted with the meaning of the word;
5) morphological analysis of the word, study of its division into syllables and listening to its pronunciation;
6) get a list of words that can be combined with the search word on the right and left (N-gram technology);
7) know the frequency or statistics of word usage;
8) determine the relationship of form and meaning of the word (homonyms, paronyms, antonyms, synonyms)

These possibilities are reflected in the corners of the corpus. As an example, we see the result by typing the word "in our lives" in the search bar in the "Dictionaries" section of the corpus:

1) "word structure" - as a result of automatic morphological analysis it becomes clear that the basis of the word "in our life" is "life";
2) "Word syllables" are highlighted;
3) an explanation of the word life is provided;
4) there are no contradictory forms;
5) the meanings of the lemma of life are presented: alive, life, life;
6) it is found that there is no paronym.

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES 2(12): 161-166, December 2021
DOI: https://doi.org/10.37547/philological-crjps-02-12-31
ISSN 2767-3758

Crossref doi Google Scholar
Accepted21th December, 2021 & Published 26th December, 2021

Picture 4. The results of the search for "in our lives" in the "Dictionaries" section of the Uzbek language educational corpus

As can be seen from this example, the educational building includes a synonymizer program [9], which provides a non-morphological, lemma-based meaning from a database of synonyms, independent of other resources [10].
The size of the educational corpus. Currently, the volume of linguistic data of the Uzbek language educational corpus has the following indicators:

№	The name of the base unit	number
1	Books	129
2	Internet texts	101 835
3	Tagged words	42 320
4	Sentences	1 221 769
5	Contexts	1 147 658
6	The number of interpreted wors	3260
7	Synonyms	993
8	Antonyms	870
9	Paronym pairs	558

CONCLUSION

1. Language corpus - a system with electronic search capabilities to determine the characteristics of national language units, a set of digitized written and spoken texts of the natural language.
2. It will be widely used by linguists,

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES 2(12): 161-166, December 2021
DOI: https://doi.org/10.37547/philological-crjps-02-12-31
ISSN 2767-3758

Crossref doi Google Scholar
Accepted21th December, 2021 & Published 26th December, 2021

lexicographers, computer linguists, programmers, editors, translators, journalists, publishers, scientists, teachers, students and any other professional.

3. It is advisable to rely on the corpus, where the text database is regularly updated, so that language education is not interrupted.

REFERENCES

1. National corpus of the Uzbek language – an important cultural reality (interview) / "Ma'rifat", 11.08.2021. № 32 (9357).
2. Lyashevskaya O. N., Sharov S. A. Frequency dictionary of modern Russian language (on the materials of the National corpus of Russian language). M.: Azbukovnik, 2009. (electronic book)
3. Grishina E. A., Lyashevskaya O. N. "Grammar dictionary of new words of the Russian language". http://dict.ruslang.ru/gram.php
4. Lyashevskaya O. N., Sharov S. A." New frequency dictionary of Russian lexicon". http://dict.ruslang.ru/freq.php
5. Kustova G. I. Combinations of words with a high degree of meaning. http://dict.ruslang.ru/magn.php
6. Biryuk O. L., Gusev V. Yu., Kalinina E. Yu. Dictionary of verbal collocations of non-objective names of the Russian language. http://dict.ruslang.ru/abstr_noun.php
7. Mixaylov M.N. Computer support of the text corpus (user's view) / M.N. Mixaylov // Rusistika segodnya. – 1998. – № 1-2. – S. 192–202.
8. Abjalova M.A. Possibilities of lexicographic search of words in the national corpus of Uzbek language. // Computer Linguistics: Problems, Solutions, Prospects / Proceedings of the Republican Scientific and Technical Conference. Electronic publication / ebook. – Tashkent: TSUULL, 23.04.2021. – B.12-17.
9. https://xn----7sbbaqhlkm9ah9aiq.net/news-new/sinonimizator-ru.html
10. Abjalova M. The issue of creating a synonymizator or synonymizer in the National Corpus of the Uzbek language // Theoretical and practical issues of creating the Uzbek national and educational corpus / Proceedings of the International scientific-practical conference. May 7, 2021. Electronic publication / ebook. –Tashkent: TSUULL, 2021. – 330 p. – B. 38-40