Advances in Economics, Business and Management Research, volume 81
1st International Scientific Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth" (MTDE 2019)

The issues on the optimal structure of lexicography block in national language and processing of its software

Mammadova R.H.
Institute of Linguistics named after I.Nasimi
of National Academy of Sciences of Azerbaijan
Azerbaijan, Baku city
rena.memmedova.1991@inbox.ru

Guluzade S.F.
Sumgait State University
Azerbaijan, Sumgait city
winslet-kate@list.ru

Aliyeva N.A.
Institute of Linguistics named after I.Nasimi
of National Academy of Sciences of Azerbaijan
Azerbaijan, Baku city
rajabova.narmin@gmail.com

***Abstract* — The article is devoted to the work done in the preparation of blocks of the Azerbaijan National language corpus, and the software used in the preparation of this lexicography. Since the problems of national languages corpus are one of the newest in recent years, the materials and research in this article are of particular important. The article describes the content of the national-language corpus, the description and comparative descriptive methods were used during the research. The main goal of the article is to determine the optimal structure of lexicography in national linguistic corpuses by analyzing the existing modern computational linguistic systems. At the same time, the study of a computational lexicography and the using of its software provide the scientific novelty of the article. The materials presented in this study are particular scientific important in terms of the construction of language corpus. Because the studies in this direction are impossible without special software. For this purpose, the article analyzes software that can be used to develop the corpuses of the national languages of the Azerbaijani language. At the end of the article, the results of the study were summarized and noted that it is important to use the existing methods and tools of information technology in the formation of the optimal structure of lexicography blocks and its software using.**

Keywords* — *national language corpus, lexicographies blocks, computational linguistics, machine translation, software.

I. INTRODUCTION

The period that we live is called century of information and high technologies. Now, to learn about events in the world, to get acquainted with the problem, is applied to more electronic libraries and scientific bases. From this point of view, the creation of national linguistic corpuses is an urgent problem. For the first time in our country, it is planned to create a national corpus of the Azerbaijani language, to prepare its optimal structure of lexicography blocks and corresponding software. The research results can be used in scientific-practical and technical systems to create national corpus for the Azerbaijani language, as well as other Turkic languages. In this direction, the block of lexicography within the structure of the national language, its optimal structure and the preparation of modern software are important.

II. THE ESSENCE AND CONTENT OF AZERBAIJAN NATIONAL LANGUAGE CORPUS

National language corpus refers to a search engine, designed to collect texts in electronic forms in any particular language. National corpus of a particular language, the language is fully represented in all stages, genres, literary language, and dialects at certain stages of the historical period [3]. However, the texts in the corpus are not in irregular, they are arranged in a certain rule and order. Full-text products in electronic forms are placed in such a way that the user can get definite information if necessary. That is, the corpuses of national languages have other features than electronic libraries and scientific bases. The national corpus is not a collection of “exciting”, more “readable” works, like electronic libraries. The national corpus is intended to represent the language completely and in all styles. Of course, at later stages it may be possible to provide a wider and fully representative of language, as well as create electronic versions of all samples of folklore, classical literature and materials in dialects. In all versions, the corpus of national anguages must be set so that the user has access to definite information.

The creation of a national corpus for specific languages was in fact previously studied on a global scale and at that time was called machine fund. Both terms, machine fund and national corpus, are used in parallel.

Some researchers have interesting ideas on this issue. They show that according to the current state of terminology in the field of information technology, it is more acceptable to consider the use of a term “computer fund” instead of a

"national corpus". This indicates that the term is still not completely stable [4].

The British National Corpus (BNC) was first created by Oxford University in the late 1980s and early 1990s and were collected about 100 million words in various genres (for example, literacy, magazine, newspaper, etc.) [9].

It should be noted that the texts included in the corpus of national language must contain all styles of the Azerbaijani language at the same level and in the same volume. The larger the cops' size more reliably the corpus is considered. Modern information technologies allow in any degree to increase the volume of corpuses of national languages. It is advisable to use the optimal placement methods here. Much depends on planning the structure of the corpus. In general, if the text written in Azerbaijani language and any text belonging to this language that present the language can be added to corpus.

III. NATIONAL CORPUS OF TURKIC LANGUAGES

In 1988, at the XIV Plenum of the Soviet Turkologists Committee in Moscow, it was decided to start work on the creation of a machine fund of the languages of the Turkic peoples inhabiting the former USSR (Azerbaijan, Uzbek, Kazakh, Kyrgyz, Turkmen, Bashkir, Tatar, Gagauz, Yakut, Chuvak, Khakas, Altai, Kumyk, Noqay, Karaim) [9]. Any information about Turkic languages should be collected and systematized here. It was decided to choose the Linguistics Institute of the Academy of Sciences of the Republic of Kazakhstan as a center for creating a machine fund of the Turkic languages.

The following information was considered necessary in the creation of the national corpus of the Turkic languages:

Structural-phonetic information covering different types of one syllable Turkic word roots;

Morpheme lists;

Schemes that reflect syntactic relationships;

The grammar thesis of affixes;

A collection of analytical indicators on the phonetic, grammatical structure of specific Turkish languages.

The linguistic bank of Turkic languages was supposed to be created taking into account the above mentioned issues [2].

Different corpuses have been created in the Turkic language belonging to the family of Turkic languages. One of them is the oral Turkish language corpus ("Sözlü Türkçe Derlemi" (STD)) supported by TÜBİTAK during 2008-2010, is an online corpus of 1 million words, obtained through interviews or various communication tools.

The Turkish Language Corps (TUD) users can search by using the information covering the period of 1990-2010 years "books, periodicals, different types of textual and non-published text", from nine different fields, including social sciences, arts, trade and finance, thinking and belief, world problems, applied sciences, natural and fundamental sciences; "the author's gender (male, female), author, types of authors (single, multiple, organization), readership (children, youth, etc.)".

As a result of search of the word "toy" (oyuncaq) in the survey interface, in the upper part of the screen was shown a total of 4458 texts where this word were used in several different texts (450), how many times was used (1200), and was given a the frequency of this usage in million words [15].

The other of Turkish online corpus — TS Corpus is a universal, unbalanced corpus that helps the user a lot by tagging the word type, morph and root word. You can enter corpus in the registry menu by creation a username and password. When you enter the word "face (yüz-üz)" in the search interface, you will get 46,497 results with (Simple query (ignore case)) the word "face (yüz-üz)" without a difference big and small letters. Looking for the difference of the big and small letters (a simple query case sensitive) of this search, the results will be got so: 41,656 with small letters, 4,413 with big letters.

If the root of the word (lemma) is written as {KÖK} in the query interface, the simple root and correction form is obtained. There are two benefits to searching so. The first of these benefits is that when you examine the word {gönül} you can find the variants of this words such as "gönlüm, gönlün" as a result of reduction of vowel during the process of adding suffixes of "-ım, ın" and others. The other benefit is getting the forms of words "b,c,d,ğ" as a result of vowel adding process to the words with "p,ç,t,k" endings.

The third online corpus- the historical derivative of the ancient Turkic and Karakhanids Turkic (ETKT-D) is the historical corpus of the ancient Turkic and Karakhanids Turkic. ETKT-D is an online corpus of Turkic art works with 400-450 thousand words, covering a 600-year period, introduced into the electronic form using a combination of words and syntax of written scripts belonging to Orkhon, Turkish Uyghur language and Turkic Karakhanids. No need to enter a username or password to register to corpus.

The latest Internet corpus is the electronic corpus of Turkish texts before Islam. Language of the site is given in German. Access to the division of the ancient Turkic words written in the alphabets Turkic Khaganate (Runik), Uyghur, Mani, Tibetan, Chinese, Syriac and Brahmi before Islam, can be obtained in the menu "Writtings of words query form".

Academic Dictionary — grammatical foundation is considered one of the most important components in the national corpus of the Kazakh language. This fund consists of subtitles. In these substructures historical and etymology, dialectology, onomastic, grammar, lexica and vocabulary are based on specific communication schemes.

As in the national corpus of any language, it is assumed to have rich software in the national corpus of the Kazakh language. This software should allow full linguistic analysis. Here must be included automatic morphological, syntactic, semantic analysis. A system that provides complete analysis is functioned as a linguistic processor.

Work on the creation of the machine fund of the Bashkir language began since 2003. Dictionary entries on 500,000 vocabulary units of the modern Bashkir language were given in the substructures lexicography of the corpus. Lexicographic substructures consist of dictionaries of academic and educational profiles — monolingual, bilingual, multilingual, frequency, terminological, phraseological, synonyms, survey dictionaries, onomastic dictionaries representing places (streets, cities, accommodation etc.).

Currently, the corpus has prepared electronic versions of 579 papers by 63 authors, covering the period from the beginning of the twentieth century to the present, and word forms were edited in a total volume of 9277754. These texts were adapted to the new orthography rules of the Bashkir language, adopted in 1981 [2, 54-58].

IV. CREATION OF MODERN NLP SYSTEMS

The work on the creation of modern NLP systems (processing of national languages) of Turkic languages has recently become more urgent. It can also be observed according to information on the Internet. Original NLP systems were created in various Turkic languages for speech recognition, text-to-speech, machine translation and the creation of independent search engines. Examples of such modern NLP systems are:

For the Turkic languages:

- SR (speech recognition) [15].
- MT (machine translation) [6].

For the Kazakh language:

- SR (speech recognition) [11].
- MT (machine translation) [10].

Studies on the creation of NLP systems in Azerbaijan are conducted as part of the "Dilmanc" project. [9]. A lot of work has been done on the project and it is planned to continue work on this field. In the same project, it is important to note bilingual corpuses that cover all language styles, and monolingual corpuses for specific languages. These corpuses and their dimensions are shown below:

English-Azerbaijani bilingual corpus – 2 million sentences.

Turkish-Azerbaijani bilingual corpus – 277 thousand sentences.

Russian-Azerbaijani bilingual corpus – 4.5 million sentences.

Azerbaijani monolingual corpus – 60 million sentences.

Turkish monolingual corpus – 322 million sentences.

V. COMPUTATIONAL LEXICOGRAPHIC AND PROCESSING ITS SOFTWARE

Computational linguistics contains the object and subject of corpus linguistics. O.S.Rublyova relates the studies in national corpus field of the Russian language to the subject of computational linguistics. Lexicographic material on computers is stored and used in machine card files. With the development and improvement of computer technology, more perfect lexicographic database has been created. According to O.S.Rublyova, the "national corpus of the Russian language" can be considered as one of such lexicographical foundations [16]. Since then, the electronicization of dictionaries has become extensive.

The corpus- stored on an electronic carrier is a set of natural language texts set in a certain order providing material for linguistic studies related to different language events and aspects in the world linguistics, including Russian linguistics. The texts that contain the corpus and to apply them are based on certain rules. The first multifunctional databases and lexicographic were created in the USA in 1956. Numerous lexicographic have been processed based on electronic databases. Webster's English dictionary can be exampled to this [17]. The first corpuses were also formed during this period (The Brown Corpus, Lancaster-Oslo/Bergen Corpus, London-Lund Corpus). However, the number of words in these corpuses was limited.

Computational lexicographic should be studied as a field of computational linguistics in that context. This term also includes a term statistical lexicography that was trendy and highly by the time.

Computational lexicographic differs from ordinary lexicographic in the following characteristics:

Application to the lexicographic and obtaining information based on certain queries easier and faster.

It can be applied to several lexicographic at the same time and quickly determined the meaning of a word and possible to get a general conclusion.

Each of these lexicographic has the ability to expose any language and compare it with other lexicographic.

Ordinary lexicographic stay unchanged till reprints, the number and description of words can't be changed. Computational lexicographic are open, dynamic systems. It can be added new words and delete archaic words in such lexicographic.

The history of Computational lexicographic in Azerbaijan begins with the 70s of the last century. For the first time over the years, the frequency lexicographic of newspapers in Azerbaijan was compiled using electronic computers. In subsequent years, much work was done on the processing of software in Computational linguistics, as well as Computational lexicographic. Currently, these works are being developed using modern software engineering mechanisms and technologies [18-19].

VI. AZERBAIJANI LEXICOGRAPHIC AND MACHINE TRANSLATION SYSTEMS

These bilingual and monolingual corpuses are successfully used in the systems of formal linguistic analysis and machine translation of the "Dilmanc" project. Bilingual parallel text corpuses are used as an automatic translation and lexicographic between the Azerbaijani, Turkish, Russian and English languages. Monolingual corpuses are important for checking the accuracy of translated texts.

The following electronic dictionaries are also available in the bibliography of Azerbaijani lexicographic on the Internet:

1. AzerDict — the largest Azerbaijani free online Azerbaijani-English, English-Azerbaijani dictionary;

2. Intelsoft — translation system from Russian to Azerbaijani.

3. Google Translate — English-Azerbaijani, Azerbaijani-English machine translation.

4. Azerbaijani-Turkish dictionary and others.

All these electronic dictionaries and machine translation systems can be considered components of the national corpus of the Azerbaijani language that will be created in future.

VII. CONCLUSION

The investigations and analyzes show that the creation of the modern Azerbaijani national language block is an actual issue.

The processing of a lexicographic block within the national language corpus is of great importance as part of Computational lexicographic direction.

It should be used effectively with the capabilities of modern ICT, including software engineering elements, technology and mechanisms, in the formation of the optimal structure of lexicographic blocks and its software usage.

References

[1] L.A.Buskunbaeva, Z.A.Sirazetdinov, “On the problems of the national corpus of the Bashkir language”, Materials “Modern Kazakh linguistics: topical issues of applied linguistics”. Almaty, 2012, p. 54-55.

[2] U.Y.Sharapova, G.N.Babshanova, “Computational lexicography as one of the directions of modern applied linguistics”, Collection: Actual problems of linguistics. 2013. Materials of the All-Russian scientific-practical conference of students, graduate students and young scientists, 2013, p. 176-179.

[3] Mahmudov M. "Computational linguistics", Baku, "Science and education", 2013, 356 p.

[4] Mahmudov M., Fatullayev R. and others. "Theoretical and practical issues of creation of NLP systems and national corpuses for the Azerbaijani language", Turkology, 2016, № 4, Baku, p. 15-28.

[5] Mahmudov M., Mammadova R., "Problems of mathematical-statistical methods of the Azerbaijani language and new technological approaches: problems, perspectives", №1, Institute of Linguistics named after Nasimi, Baku, 2016, p.18-29.

[6] http://cevirsozluk.com (Date of applicaton: 20.11.2018)

[7] http://corpus.byu.edu/bnc (Date of applicaton: 09.01.2019)

[8] http://dergi.kmu.edu.tr/userfiles/file/Mayis20142/30m.pdf (Date of applicaton: 11.09.2018)

[9] http://dilmanc.az/ (Date of applicaton: 07.02.2019)

https://sozdik.kz/ru/dictionary/translate (Date of applicaton: 07.02.2019)

[11] http://uniline.kz/wordpress/?p=665 (Date of applicaton: 20.11.2018)

[12] http://www.dam.org.tr/index.php/tr/derlemler/66-soezlue-tuerkce-derlemi (Date of applicaton: 23.12.2018)

[13] http://www.natcorp.ox.ac.uk/ (Date of applicaton: 22.12.2018)

[14] http://www.turcologica.org/rossijskij-komitet-turkologov (Date of applicaton: 20.11.2018)

[15] http://www.sestek.com/tr/konusma-tanima (Date of applicaton: 20.11.2018)

[16] http://www.ruscorpora.ru/ (Date of applicaton: 10.02.2019)

[17] http://www.webster-dictionary.org/ (Date of applicaton: 03.03.2019)

[18] Veliyeva K.A. “Modern trends in computational linguistics. Information Society Problems”, 2016, №2, p.98-107.

[19] Aliquliyev R.M., Qurbanova A.M., Terminology Informatics: Stages of formation and development trends. Express information. Information society series. Baku: "Information Technologies" publishing house, 2014, 71 p.

[20] Aliquliyev R.M., Qurbanova A.M., "Conceptual bases of creation of terminological information system in Azerbaijan", Information society problems,Baku, 2011, №1, p. 3-8.