Problems of Creation of the All-Turkic National Corpus

Gulzhan Doszhan

L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan

Abstract

The paper presents the results of research on the theoretical and practical issues of the creation of national corpus of the Turkic world.

This paper consists of four parts and conclusion. The first section is devoted to a theoretical analysis of a problem, the second section describes a brief history about the ideas of creation of the International machine fund of Turkic languages from the former USSR, the third section considers the realization of concerted reforms of the Turkic countries on creation of the all-Turkic terminological fund and the fourth section analyzes the significance of future projects of the Turkic people on corpus linguistics.

Key words: national corpus, corpus linguistics, information technologies, Turkic world, globalization, all-Turkic terminological fund.

1. Introduction

Corpus is defined as a collection of texts stored in an electronic database. A corpus represents varieties of spoken or written text types, sampling language in use, providing researchers the most fundamental database upon which they can search various aspects of language. Rapid developments in the recent computer technology made it possible to access and process huge amount of data and consequently have caused a shift in linguistic research from introspective data to naturally occurring data. This shift ultimately led to the construction and use of a number of well constructed language corpora which in turn help recognize the importance of real data representing those aspects of language that were not possible to observe otherwise. Furthermore, corpus construction processes contributed enormously to the emergence of a new field in linguistics, namely corpus linguistics and also to the efficiency of the applications of various natural language processing studies.

In the last three decades, 190 different corpora representing 80 different languages have been constructed. Today, most of these corpora are organized under various international consortiums and are available for users for a variety of purposes. Such corpora serve not only for general and applied linguists but also computational linguists and educational linguists alike.[1]

Corpora of texts are used typically to study the structure and function of language. The distributions of various linguistic units, comprising texts in a corpus are used to make and test hypotheses relevant to different linguistic levels of description.[2]

Summarizing the aforesaid definitions it is possible to claim that a corpus is a reference system based on an electronic collection of texts composed in a certain language.

The priority direction of modern applied linguistics is the corpus linguistics. The corpus linguistics defines the general principles of creation of linguistic corpuses of data (corpuses of texts) with the use of modern computer technologies, develops a technique of collecting the real language phenomena of texts of written and oral speech, and also ways of their storage and the analysis. Work with corpuses allows to abstract in a certain degree from subjectivity of the researcher and to come nearer to objective studying of language. A national corpus represents wide information about the dynamic development of a definite language in all the variety of genres, styles, territorial and social variants of usage, etc.

A national corpus is distinguished by two features. Firstly, it is characterized by representative and well-balanced collections of texts. This means that such a corpus contains, if possible, all the types of written and oral texts presented in the language (various genres of fiction, journalistic, academic, and business, as well as dialectal and sociolectal, texts). The proportion of text types in the corpus is based on their share in real-life usage at the time of composition. A representative corpus usually contain up several million tokens.

Secondly, a corpus contains additional information on the properties of texts. It is achieved by means of annotation which is a principal feature of the corpus, distinguishing the corpus from simple collections (also known as «libraries») of texts on the Internet, such as, in Russian, the Maksim Moshkov library or the Russian Virtual Library. Such libraries are not well suited to academic work on the nature of language; they tend to focus on the content of texts rather than their language properties, while the creators of the Corpus recognize the importance of literary or scientific value of the texts, but see them as a secondary feature .[3]

A national corpus is being created by linguists and IT-specialists for academic research and language teaching. Most of the major world languages have their own corpora. A well-recognized example is the British National Corpus, which is used as a model for many modern corpora. Among the Slavic languages, the Czech National Corpus (compiled at the Charles University of Prague) is more notable, and for the Turkic languages, the Turkish National Corpus (designed by a team of linguists from Mersin University and is funded by the Scientific and Technological Research Council of Turkey) is advanced model.

The main purpose of the corpus is to facilitate academic research on the lexicon and grammar of a language, as well as the subtle but constant processes of language change within a relatively short period of time: from one to two centuries. The other purpose of the corpus is to serve as a reference point for lexical, grammatical, and accentological questions, and the history of the language. Modern IT-technologies make the processing of large volumes of text significantly simpler and faster, which creates the possibility for mass statistical analysis of texts. As a result, language research now yields results which could only be guessed at previously. Nowadays, truly scientific descriptions of grammars and academic dictionaries must be based on corpora of their respective languages. The use of corpus data is desirable (if not always strictly necessary) in other, more specialized language research.

Therefore, the main users of national corpora are linguists of various profiles. Nevertheless, the corpus is useful for non-linguists too. Reliable statistical information on language use in a certain period or by a certain author may be of interest for researchers of literature, history and other humanitarian subjects. National corpora are also useful for

language teachers, both native and foreign; language textbooks and teaching programs are increasingly oriented toward corpora. A corpus can be used for ascertaining the variants of usage of unknown words by foreigners, students, teachers, journalists, writers. Therefore, the corpus is aimed at people who are interested in the structure and usage of a language .[3]

2. The Idea of Creation of the International Machine Fund of Turkic Languages from the Former USSR

In 1988 at an extraordinary meeting of plenum of All-Union committee of turkologists in Moscow the idea of creation of the International machine fund of Turkic languages (MFTL) was put forward and supported. As a result the structure of the united working group included known scientists from St. Petersburg, Moscow, Novosibirsk, Baku, Tashkent, Bishkek, Kazan, Ashkhabad, Ufa, Nalchik, Cheboksary and Almaty. Plenum adopted the relevant resolution. The essence of fund consisted in creation of global worldwide network of the server bases connected together with the uniform program of input of linguistic data on all living and dead Turkic languages. It was supposed that the machine has to give out, according to inquiry, any morphological, syntactic, phonetic-phonologic or lexico-semantic data of the synchronous or diachronic contents, whether it is the textbook or the dictionary of concrete language or language of its related group, etc.

The initiative was supported by well known scientists from Leningrad and Moscow, and also from other regions of the USSR. At that period of time, in Almaty under the leadership of academician A. Kaydarov the powerful turkological school with divisions on mathematical linguistics, a historical lexicology and reconstruction of classic languages was created. Unfortunately, with the collapsing of the Soviet Union, this grandiose project was stopped.

3. The Current Reforms of Creation of All-Turkic Terminological Fund

In the present time the creation of all-Turkic national corpus is erected in a rank of important historical-cultural, informational and political goals of the Turkic World. Therefore a research of a problem of creation of the all-Turkic national corpus is very topical in the conditions of globalization. In this regard, research of any aspect of language, including drawing up multivolume dictionaries, grammatical and lexical researches demands work with extensive arrays of texts. It is rather difficult process demanding not only intellectual work, but also time, especially at preparatory stages. Therefore it is necessary to transit modern methods of collecting a material and its analysis which will significantly increase labor productivity and will open a way to innovative methods of research of Turkic languages.

The creation of the National corpus of Turkic languages intensify a communication of fraternal peoples in a science and, being based on the international experience of cooperation of related languages, to develop strategy of assimilation of loan words and terms, to create the all-Turkic terminological fund by means of modern information technologies. The similarity of Turkic languages which have much in common both with grammatical system and lexical structure give an opportunity to create an all-Turkic national corpus.

In June 1924, during the First Congress of Kazakh Intelligency which congregated in Orenburg, the founder of the Kazakh linguistics, great scientist

Ahmet Baitursynov has proved how the words of Turkic people could be used and noted that: «In absence of the necessary terms in Kazakh, they should be borrowed from the languages kindred to Kazakh. It is performed on the following grounds:

1. although the most of words of the kindred languages do not have the common forms but have the common roots, so they are easily understood, heard and they are not as strange for pronunciation as a word of non-kindred language;

2. Turkic people had and have the continiuos communication among themselves, and therefore the most of the words of one language can be known for the representatives of another language without any common roots.[4]

After the long time, the first attempts to facilitate in collaboration of Turkic-language countries in this regard were taken in 1999 when a special task group was established through the help of the Turkish Information Society. October 2001, the First Turcological Forum was arranged with regard to the collaboration in information technologies. Later on, the meetings in Azerbaijan, Turkmenistan and Kyrgyzstan were arranged with regard to various spheres of terminology. In 2011 the 9th Forum of Terminologists of Turkic Countries was arranged in Astana by the Committee on Languages of the Ministry of Culture of the Republic of Kazakhstan together with the Turkey Committee on Languages and Turkish Society of Information Technology. The turkologists gathered at the forum to try and strengthen ties of fraternal peoples in science and, basing on worldwide experience of cooperation of kindred languages, develop a strategy for borrowing and unifying terminology; create a common fund of industry terminology, especially in information technology.

In recent years the idea of the creating of the common Turkic terminological fund gained a new impulse. To it promoted establishment Cooperation Council of Turkic Speaking States (CCTS) in 2009 as an international intergovernmental organization, with the overarching aim of promoting comprehensive cooperation among Turkic States. Its four founding member States are Azerbaijan, Kazakhstan, Kyrgyzstan and Turkey. Turkmenistan and Uzbekistan abstained from accession to this organization.

In May 25, 2010 in Astana President of Kazakhstan Nursultan Nazarbayev and President of Turkey Abdullah Gül opened the new research center – Turkic academy. The initiative of establishment of Turkic academy which would be engaged in studying and research of language, history and culture of the Turkic people, belongs to the Kazakhstan leader and to them for the first time was stated in October, 2009 at the IX summit of Heads of the Turkic countries in Nakhichevan (Azerbaijan).

In consideration of the recommendations by the Council of Wise Men, which serves as the advisory board of the CCTS, terminology committee was set up with the participation of academics from the member states of this organization in 2012. First Meeting of the Terminology Committee, founded with a view of convergence among national languages of the Turkic Council, was held in Istanbul on November 16, 2012.The Meeting brought together scholars commissioned by the Governments as national representatives as well as experts from member states of Turkic council, Turkic academy, heads and analysts of Turkic linguistic structures.

Participants elaborated on the basic principles of developing common terminology and agreed that the related academic endeavors should be collected

under a single roof and expedited. Other issues agreed upon during the Meeting include preparation of a glossary of common terms and an illustrated explanatory dictionary of common words as well as further improvement of the Comparative Dictionary of Turkic Languages. It is expected that the Committee will convene several times a year and the organizational actions in the sphere of all-Turkic terminology will be carried out by the Turkic Academy.

4. Relevance of Creation of the All-Turkic National Corpus in the Conditions of Globalization

The end of XX and beginning of XXI centuries were marked in the world by dramatically intensified globalization process. World community experiences a complex stage of all social processes which in particular is conditioned also by the development of information technologies. Globalization covered economical, political and cultural spheres of the society. This phenomenon is discussed in all branches of modern science: sociology, culturology, political science, and, of course, in linguistics.

In this connection, K. Khanazarov, a well-known scientist on linguistic philosophy, considers that «Globalization is an objective process which by no means is aimed at causing damage to the existing languages. However, it breaks the basis of languages by its speed-up and expansion, destroys the foundation on which thousands of languages are based especially languages of small nations, folks, tribes and ethnic groups».[5]

In the present time, the development of computer programs, entry of the language into the Internet space is one of the important tasks for linguists and programmers of Turkic states.

Nowadays, in the connection with absence of a uniform standard alphabet, the Turkic people are divided behind frameworks of interstate sphere into spheres of action of other regional languages – Russian, English, Farsi, Chinese etc. For official mutual dialogue, for example, between the Turkic CIS countries Russian, between Turkey and the Turkic CIS countries is used English, Turkic from Iran and Turkic from Afghanistan most likely will use Farsi for dialogue among themselves and etc.

Unfortunately, the all-Turkic national corpus in force a number of objective and subjective factors still isn't developed. At present among 30 Turkic languages only the Turkish language has the own national corpus (http://www.tnc.org.tr/index.php/tr/). Nevertheless, some Turkic languages have the electronic portal or machine fund, such as: Bashkir (http://mfbl.ru/), Tatar (http://klf.ksu.ru), Kazakh (http://til.gov.kz), Tuva (http://tuvancorpus.ru) and Shor (http://shoriya.ngpi.rdtc.ru/). Besides, together with the German scientists corpuses on monuments of Turkic languages, such as the corpus on monuments of the runic letter of Mountain Altai (http://www.altay.uni-frankfurt.de) and the electronic corpus on monuments of pre-Islamic classical Turkic texts (http://vatec2.fkidg1.uni-frankfurt.de) are developed.

5. Conclusion:

In creation of the all-Turkic national corpus there are vital issues, the decisions which depend on joint efforts from the Turkic states:

1. The main barrier in creation of national corpus of the Turkic world is the differences in graphic basis of their alphabets. The most of Turkic people use three different graphic systems of the letter: Latin, Cyrillic and Arabic.

Necessity of a uniform graphic basis of alphabets is obvious.

The majority of the independent Turkic countries (Turkey, Azerbaijan, Uzbekistan and Turkmenistan) use Latin as a graphic basis of the national alphabet, only Kazakhstan and Kyrgyzstan still use Cyrillic. Therefore, first of all, it is necessary to create the all-Turkic alphabet on the basis of Latin graphics. The alphabets of Turkic languages which are based on Cyrillic are not functioned in full on software of modern computers.

Modern development of the Internet and concerted plans of the Turkic states in the cultural and humanitarian, political and economic spheres allows saying that creations of the all-Turkic national corpus in the future will be a guarantee of a long integration between the Turkic people and preservation of all Turkic languages and cultures in the conditions of globalization.

2. The absence of a special joint program (memorandum) on creation of the all-Turkic national corpus.

3. The absence of the mechanism of stimulation of scientists engaging in research on corpus linguistics.

4. Low level of institutional development of corpus linguistics in the Turkic countries.

Observing the foreign experience, each Turkic country developing their own national corpus can start the collaboration over creation of the national corpus of the Turkic world using the integrated system of information technologies in the field of linguistics.

References:

[1] Aksan, Yeşim ve Mustafa Aksan, «Building a national corpus of Turkish: Design and implementation», *Working Papers in corpus-based linguistics and language education, Tokyo: Tokyo University of Foreign Studies*, pp. 299-310, 2009.

[2] K. Ahmad, D. Cheng, T. Taskaya, S. Ahmad, L. Gillam, P. Manomaisupat, H. Traboulsi, and A. Hippisley. "The mood of the (financial) markets: In a corpus of words and of pictures", Corpus *Linguistics Around the World. Ed. Andrew Wilson et al. Amsterdam: Rodopi Publishers*, pp. 17-32, 2006.

[3] http://www.ruscorpora.ru/en/corpora-intro.html

[4] Kurmanbayuly, Sherubay. «It is necessary to create common terminological fund for related languages», *Materials of scientific-practical conference «Problems of expansion of the sphere of use of a state language», Kokshetau: Kokshe-polygrafia.*, p. 30, 2001.

[5] Khanazarov, Kuchkar. « Problems of development of philosophy of language», Tashkent: «Uzbekistan», p.134, 2007.