THE CORPORA OF THE BASHKIR LANGUAGE

Z.A.Sirazitdinov
Institute of History, Languge an Literature,
of Ufa scientific Centure
Russian Academy of Sciences
sazin11@mail.ru

ABSTRACT

The article discusses the state of corps linguistics in the domestic and foreign linguistics and design issues of corps at the Institute History, Language and Literature, Ufa Research Center. The author analyzes the work of the laboratory of linguistics and information technology in this area. We describe the proposed methods of creating of corpora of the Bashkir language, analyzes the results obtained, the task for the future.
Keywords: corpus linguistics, the Bashkir language, information systems, applied lingstistics.

Keywords: Corpora, Bashkir language, Database.

1 INTRODUCTION

Originated in the second half of the 20th century trend in foreign linguistics associated with computer processing of large volumes of text, has formed a new rapidly growing trend of Philology - corpus linguistics.

Now all the major languages have got their national corpuses. In Russia are being developed corpuses of languages of indigenous peoples: Buryat [1], Kalmyk [2], Lezgian [3] Ossetian [4], and others. Scientific developments and corpus projects to the Turkic group of languages It should be noted here: Kazakh [5], Tartar [6], Tuvan [7] Turkish [8], Shor [9], Khakassian [10], Kumyk [11] Kyrgyz [12].

Given the urgency of corpus development, the laboratory of Linguistics and Information Technology of the Institute History, Language and Literature (RIHLL) of the Ufa Research Center the Bashkir language corpususes are being developed in two ways: a) the corpus of prose texts; b) the corpus of journalistic texts.

In line with the above directions in RIHLL of he URC RAS we have developed an integrated system that allows you to create corpora, carry a wide range of search tasks and maintain enclosures. Our design implementation is based on the Oracle DBMS. It is essentially a second direction in the national corpus technology (after software development of the Yandex company for corpus of Russian language).

2 THE CORPOPA OF BASHKIR LANGUAGE

The integrated system consists of two parts: the user and administrator ones.

1. The user unit includes the following software:

a. Means of determining the volume of the corpus wich allows to allocate the user subcorpus.

b. Searching funds. There are programs allowing a flexible search for such linguistic parameters as word form, lemma, semantics, grammatical categories of inflection, grammatical categories, the combination of grammatical categories, the combinations of grammatical subcategories, the combinations of word forms, combinations of lemmas.

c. Programs of quantitative-statistical analysis of the text body. These tools are now under development. Today are being developed functions of plotting frequency dictionaries of word forms and tokens.

2. The Administrator Unit (with rights of entry for laboratory staff). Which Includes the following softwares:

a. Software for input and automatic text markup. These tools produce morphological and semantic markup of new texts introduced.

b. Editing tools. It is possible to edit the main dictionary, lists of inflectional categories, patterns and source text.

c. Tools for resolving the grammatical and lexical homonymy in the interactive mode. In that case laboratory staff can view the text and eliminate homonymous phenomena that cannot be resolved by the system.

d. Equity decisions on "non Bashkir words". In the process of morphological analysis of word forms of texts system one is confronted with situations where there is no identification with basic vocabulary, or with list of inflection affixes. These word refer to "non Bashkir words" of vocabulary. "Non Bashkir words" consist of neologisms, dialect words and blotches from other languages. The Software allows to correct typos, add new basis or markup a word form as blotches. Languages sources of foreign vocabulary can be added or removed from the relevant list.

e. Programs of statistical records of the work of server with corpora.

f. Programs of export any markup text from the Oracle database in a XML format for data exchange with other national Freestanding projects.

Morphological tagging system of Bashkir corpora is focused on representation of all regular inflections of grammatical forms not always reflected and coinciding with the forms adopted in the Academic grammar. Morphological information of Bashkir word forms in the corpuses comprises: a) part of speech characteristics; b) the set of morphological features by the type of agglutinative affixes of inflection which are divided into nominal and verbal forms.

They distinguish in Bashkir 12 parts of speech: nouns, numerals, adjectives, adverbs, verbs, pronouns, imitative words, interjections, modal words, conjunctions, particles, postpositions. These characteristics are specified in the dictionary bases.

For nominal parts of speech 15 morphological categories are distinguished [13]. Verb morphological features include 11 categories [14].

The morphological analyzer analyzes the text word forms and performs the markup the text. Automatic stemming algorithm is performed on the basis of the isolation of the serial letters and word forms by comparing the remainder of the isolation of a fragment with dictionaries of bases words and affixes of the Bashkir language. Note that some languages are offered analyzers based on the dictionary of word forms that represents the list of all possible word forms of the language indicating the basics and grammatical features. In our opinion, for agglutinative languages with a slender morphological system this approach is costly, requires more system resources and a large handlabour on a divisional basis, and tie-dye grammatical categories of word forms.

For proper identification of bases words and affixes they use grammatical filters in a morfoanalizer.

1. Matching filter of phonetic structure of affixes to the phonetic structure of the base words.

For this filter, any Bashkir base word is represented as a pseudotensor element ${}^{pt}a_i$ – i- of a dictionary word, where the

p, t – determine the phonetic structure of the word, take values 1,2,3,4.

All affixes of inflection are well indexed, respectively o (p, t) to base or word forms which they can join. Table 1 illustrates a view in the database of the corpora of plural affixes.

TABLE 1. PRESENTATION OF PLURAL AFFIXES

$^{11}b^{1}_{1}$=ләр	$^{21}b^{1}_{1}$=лар	$^{31}b^{1}_{1}$=ләр	$^{41}b^{1}_{1}$=лар
$^{12}b^{1}_{1}$=дәр	$^{22}b^{1}_{1}$=дар	$^{32}b^{1}_{1}$=дәр	$^{42}b^{1}_{1}$=дар
$^{13}b^{1}_{1}$=ҙәр	$^{23}b^{1}_{1}$=ҙар	$^{33}b^{1}_{1}$=ҙәр	$^{43}b^{1}_{1}$=ҙар
$^{14}b^{1}_{1}$=тәр	$^{24}b^{1}_{1}$=тар	$^{34}b^{1}_{1}$=тәр	$^{44}b^{1}_{1}$=тар

2. Matching filter of affixes regulating rules combinations. This filter is based on the list of possible combinations of models inflectional affixes of the Bashkir language, which we have previously discussed in one of our works [15].

For this filter we compiled structural models implemented in the language type of word forms in the form of pseudotensor elements ${}_{j1k1l1m1n1v1u1x1y1z1}{}^{i}A^{jklmnvuxyz}={}^{pt}a^{f}_{i}\otimes{}^{pt}b_{j}^{j1}\otimes{}^{pt}b_{k}^{k1}\otimes{}^{pt}b_{l}^{l1}\otimes{}^{pt}b_{m}^{m1}\otimes{}^{pt}b_{n}^{n1}\otimes{}^{pt}b_{v}^{v1}\otimes{}^{pt}b_{u}^{u1}\otimes{}^{pt}b_{x}^{x1}\otimes{}^{pt}b_{y}^{y1}\otimes{}^{pt}b_{z}^{z1}$,.

So ${}_{1k1}{}^{i}A^{12}={}^{pt}a^{f}_{i}\otimes{}^{pt}b_{1}^{1}\otimes{}^{pt}b_{2}^{k1}$ determines inflection formed with plural affixes and case system. Same element ${}_{11}{}^{i}A^{12}={}^{pt}a^{f}_{i}\otimes{}^{pt}b_{1}^{1}\otimes{}^{pt}b_{2}^{1}$ defines all inflection affixes of plural genitive.

3. Filter of graphic transmission at junctions of phoneme.

Graphic changes and conversion, the consonant phonemes lossing (and some others are verified in exceptions filter).

This filter, in general, is of complicated structure and not uniform, and it has initial data base. So for some sections of the filter data are presented as indexes in the bases table.

The dictionary of word bases are the parts of speech, types of synharmonism violations and possible residues bases in inflectional processes and other options. Table 2 shows a fragment of dictionary bases.

TABLE 2. FRAGMENT OF DICTIONARY WORD BASES

И	анонс	1	
И	ансамбль	2	ансамбл
И	ант		
И	антагонизм	1	
И	антагонист	1	антагонис

3 EXPERIMENTS

The proposed principles are implemented for the corpus projects. Today the corpus of prose texts of the Bashkir language includes 997 texts from more than 80 authors. The volume of the this corpus makes up about 20 million word forms. The project is available in an online mode to the address [http://mfbl.ru/bashkorp/korpusp]. Now the debugging and optimization of corps is proceeding, work is underway to digitize the new texts.

For the corpus of newspaper texts we have prepared texts of national newspapers and magazines with a total of 5 million word forms. Extralinguistic markup system of the journalistic corpus includes the name of the press, item year, month and day of publication, article title, author. All texts are marked by category and genre. The project is available in an online mode to the address [http://mfbl.ru/bashkorp/korpuspub]. For the discussed corpus the following themes and genres are identified:

By the Subject: political and social life (politics, law, philosophy); economy (manufacturing, construction, business, finance, commerce); agriculture; art, culture and literature; science and technology; education; nature, traveling;

privacy; sports; religion; psychology; medicine; health and beauty.

Genres of texts: an interview, a conversation; articles, essays, reportage, review; advice; letters; Press review (news from other sources); congratulations; artistic and journalistic genres (essays, feuilleton, stories, poems, epigrams); review.

Today prose texts corps is being actively used by linguistics department staff in the preparation of a multi-volume Academic explanatory dictionary of the Bashkir language.

Note: English edited by Sh.Nafikof, 2014.

4 REFERENCES

[1] Badmaeva L.D., Badagarov Zh.B., Cydypov B.Z. *Obshhie problemy formirovanija korpusa burjatskogo jazyka* [Common problems of forming the corpora of the Buryat language]//Proceedings of the International Conference "Corpus linguistics – 2008". October 6-10, 2008. Sankt-Peterburg, 2008. P. 24-30. (rus).

[2] Kukanova V.V. *Arhitektura metaopisanija v Nacionalnom korpuse kalmyckogo jazyka* [Architecture of meta descriptions in the national corpora of Kalmyk language]//Bulletin of Kalmyk Institute for Humanities Research, RAS. 2011. № 1. P. 139–145. (rus).

[3] The Corpora of Lezgi URL:http://www.dag-languages.org/Lezgian Corpus/search/ (Date of circulation: 17.06.2014).

[4] The Corpora of the Ossetian language URL:http://www.ossetic-studies.org/iron-corpus /csearch/index.php?interface_language=ru. (Date of circulation: 17.06.2014).

[5] Zhubanov A.K. *Qazaq tilinin annotacijalangan matinder korpusyndagy etestik sozderge leksik-morfologijalyr belgi-kod (belgilenim) qojudyn algysharttary* [Principles of lexical and morphological marking of verbs in the Kazakh annotated corpora text]//"Tiltanym" [Linguistics]. Journal of the Institute of Linguistics named A.Baitursynov. 2012. № 1. P. 18-25. (kaz).

[6] Sulejmanov D.Sh., Hakimov B.Je., Gil'mullin R.A. *Korpus tatarskogo jazyka: konceptualnye i lingvisticheskie aspekty* [The Corpora of the Tatar language: conceptual and linguistic aspects]*]*//The Bulletin of Tatar State Humanitarian Pedagogical University. 2011. № 4(26). P. 211-216. (rus).

[7] Salchak A. Ja. *Jelektronnyj korpus tekstov tuvinskogo jazyka* [Electronic corpus of Tuvan language]//Novye issledovanija Tuvy [New Research of Tuva] (Electronic Journal). 2012. № 3. URL:http://www.new-tuva.info/journal/issue_ 15/5231-salchak.html (Date of circulation : 17.06.2014) (rus).

[8] Sözlü Türkçe Derlemi. URL: http://std.metu.edu.tr (Date of circulation: 17.06.2014).

[9] Electronic Corpora of Shor texts. URL:http:// shoriya.ngpi.rdtc.ru (Date of circulation: 17.06.2014).

[10] Shejmovich, A. V. *Morfologicheskaja razmetka korpusa hakasskogo jazyka* [Morphological tagging of Khakassian Corpora]//Rossijskaja tjurkologija [Russian turkology]. 2011 № 2(5). P. 48–61. (rus).

[11] Gadzhiahmedov N.Je. *Na puti sozdanija dialektnogo korpusa kumykskogo jazyka [*Towards creating dialectological Corporaof kumyk language*]*//Aktualnye problemy dialektologii jazykov narodov Rossii: Materialy XIII mezhdunarodnoj konferencii [Actual problems of dialectology of languages of Russia: Proceedings of XIII International Conference]. Ufa, 2013. P. 177-179. (rus).

[12] Sadykov T., Sharshembaev B. *"Manas" jeposunun ulttuk korpusun tyzyy zhonyndo [*Establishment of a national Corpora of the epic "Manas"*]*//Proceedings of the I International Conference on Computer processing of Turkic Languages (Turklang-2013). Astana, 2013. P.148-154. (kg).

[13] Sirazitdinov Z.A., Polyanin, A.D., Ibragimova, A.Sh. Ishmukhametova Korpusi bashkirskogo yazika: prinsipi razrabotki [The corpuses of the

bashkir language: design principles]//The Problems of Oriental Studies. №4, 2013. P. 65-72. (rus).

[14] Sirazitdinov Z.A., Busskunbayeva L.A., Ibragimova A.D., Ishmukhametova A.Sh. Informatsionnye sistemy I bazy dannyh bashkirskogo jazyka [Information systems and databases of the Bashkir language]. Ufa, 2013. 116 p. (rus).

[15] Sirazitdinov Z.A. *Modelirovanie grammatiki bashkirskogo jazyka*. Slovoizmenitelnaja sistema [Modeling of the Bashkir language. Iinflectional system]. Ufa: Gilem, 2006. 160 p. (rus).