THE BIMODAL CORPUS OF RUSSIAN-TURKIC BILINGUALS’ SPEECH (RuTuBiC)[[1]](#footnote-1)

Rezanova Z.I. (rezanovazi@mail.ru)

Temnikova I.G. (irtem@sibmail.com)

Nekrasova E.D. (nekrasovaed@yandex.ru)

Stepanenko A.A. (stepanenkone@mail.ru)

Datsyuk V.V. (valeriy.datsyuk@yandex.ru)

Dybo A.V. (adybo@mail.ru)

Linguistic Anthropology Laboratory, National Research Tomsk State University, Tomsk, Russia

The paper presents Russian-Turkic Bilingual Corpus (RuTuBiC) design, its basic identifying features: the aim of producing a corpus, the types of texts it contains, metatextual markup and error annotation principles, technological (IT, digital) concepts. The current state and development trends of the corpus are discussed. The corpus started as an integral part of a research project intended to explore languages and cultures’ interaction dynamics in South Siberia, it embraces the recordings of Russian-Turkic (Russian-Tatar, Russian-Shor and Russian-Khakass) bilinguals’ oral speech, transcribed and error-annotated. The corpus data allow revealing mother tongue influence within the system of deviations from the speech standard in bilingual speech by means of placing them against various sources of deviations, as well as tracing the influence of social and linguistic factors on the occurrences of deviations from the speech standard.

Keywords: Russian-Turkic, bilingual, bimodal corpus, metamarkup, error annotation, transfer.

БИМОДАЛЬНЫЙ КОРПУС УСТНОЙ РЕЧИ РУССКО-ТЮРКСКИХ БИЛИНГВОВ (RuTuBiC)

Резанова З.И. (rezanovazi@mail.ru)

Темникова И.Г. (irtem@sibmail.com)

Некрасова Е.Д. (nekrasovaed@yandex.ru)

Степаненко A.A. (stepanenkone@mail.ru)

Datsyuk V.V. (valeriy.datsyuk@yandex.ru)

Дыбо A.В. (adybo@mail.ru)

Лаборатория лингвистической антропологии, Национальный исследовательский Томский государственный университет

Статья представляет проект создания Корпуса русско-тюркской билингвальной речи Южной Сибири (RuTuBiC), характеризуются основные особенности: цель создания корпуса, типы составляющих его текстов, принципы его разметки и метаразметки, технологические решения. Характеризуется современное состояние корпуса и перспективы его развития. Корпус является частью проекта направленного на изучение динамики взаимодействия языков и культур Южной Сибири, он содержит записи устных интервью информантов, носителей русско-тюркского билингвизма (русско-татарских, русско-шорских, русско-хакасских билингвов), их расшифровку, и аннотацию речевых отклонений. Данные корпуса позволяют выявить проявления русско-тюркской интерференции в системе отклонений от речевого стандарта в речи билингвов, соотнести их с другими источниками отклонений, а также проследить влияние социальных и языковых факторов на появление отклонений от речевого стандарта.

Ключевые слова: русско-тюркский, билингв, бимодальный корпус, метаразметка, аннотирование отклонений, интерференция.

***Project overview***

The distinctive feature of the corpus is its being an integral part of the project focused on interdisciplinary research in the processes and results of cognitive and language effects caused by the languages’ interaction in Southern Siberia.

The peculiarity of the problematic field investigated within the project lies, firstly, in the fact that dynamic processes of linguistic interaction are studied not only in the aspect of changes occurring in the minor languages of Southern Siberia, but also in the processes of their influence on the Russian language, the language that is the macro-intermediary of the peoples in the region. Secondly, the cognitive and psycholinguistic substrates (grounding) of these effects are examined.

Within the project the corpus correlates with two databases (DB) that are being constructed. The first DB is a psycholinguistic one, which is a relational DB combining the psycho-linguistic, sociolinguistic data and linguistic data proper. It is intended to reveal the influence of a bilingual’s perceptual experience on forming semantics, as well as the interrelation between perceptually grounded semantics and language-formalized ways of its realization. The second DB contains the results of experimental studies focused on cognitive processing Russian language units by Russian-Turkic bilinguals. It presents the data on the subjects’ reaction variety while processing units of the lexical and syntactic level obtained in a series of directed experiments.

 Thus, the corpus materials enable the user to uncover the system of speech shifts in the speech of bilinguals under the influence of the native language, the materials of the psycholinguistic database allow revealing the aspects of the interaction between the linguistic form and the perceptual pre-linguistic experience of the speaker, the results of experiments present the data on the influence of different factors on the active processes of cognitive processing of Russian language units by Russian-Turkic bilinguals.

The DBs and corpus are integrated through the meta-markup unity: social and psycholinguistic data serve to distinguish between bilinguals’ types, which makes it possible to correlate the influence of these factors on speech practices, on the manifestation of cognitive pre-linguistic grounds of semantics formation and their effecting the interacting languages’ structures in the linguistic consciousness of a bilingual.

Thus, the corpus to be created is aimed at fixing the types of manifestations of Russian-Turkic transfer in the system of speech standard deviations in bilinguals’ speech, correlating them with other sources of deviations, sociolinguistic portrayal of bilinguals, fixing their language experience, which will make it possible not only to fix the linguistic phenomena, but also to trace the influence of social and linguistic factors on speech standard deviations’ occurring. The research is also focused on the dynamics of transfers in the Russian spontaneous speech of Russian-Turkic bilinguals, the possibility of which is achieved by analyzing the transfer phenomena in the speech of bilinguals belonging to different age groups, based on the analysis of a generalized sociolinguistic portrait of Russian-Turkic bilinguals.

The integrated data of the corpus under construction and the two DBs mentioned, intended to reveal the properties of cognitive and linguistic mechanisms of speech production and processing by Russian-Turkic bilinguals are an integral part of a project describing the dynamics of linguistic situation in Southern Siberia [Rezanova et al., 2018].

***The empirical base of the corpus***

The text database of the corpus includes the recordings of oral Russian speech of the Shors, Tatars and Khakas speaking two languages. The Russian language can occupy various functional positions, in most cases it is presented by the subjects as L2, but it is occasionally presented as L1, Shor, Tatar and Khakas being L2 (less than 1 % of all the bilinguals interviewed). The corpus includes three sub-corpora organized based on the peculiarity of language combinations ​​- Russian-Shor, Russian-Tatar, Russian-Khakass bilingualism. At present, the database of oral speech recordings of the Russian-Tatar sub-corpus includes about 350 hours of recording.

The subcorpora comprise the recordings of oral speech produced by the bilinguals with a developed language ability, aged 18 to 84. Russian-Turkic bilinguals’ speech was recorded during field studies in the areas of their dense residence (Russian-Shor – the town of Tashtagol and the neighbouring villages of Sheregesh and Bolshaya Sueta, Kemerovo region; Russian-Tatar – Tomsk and the neighbouring villages of Chernaya Rechka and Eushta, Novosibirsk and the neighbouring villages of Yurt-Ory and Akbalyk; Russian-Khakas – Abakan and the neighbouring village of Askis) as well as in the Laboratory of linguistic anthropology, Tomsk State University.

The recordings were done in accordance with the standards of the Ethical Committee of the home university, all participants signed informed consent forms, and the texts in the corpus are presented as anonymous.

Audio files with interviews are converted to text format. By now, 170 hours of recordings have been transcribed. The text is constructed keeping to spelling standards, pauses are additionally marked. Pronunciation features are recorded only in case of significant deviations from the pronunciation standards. The authors recognize the subjectivity of such approach, but it can be overcome in the practice of using the corpus, since the user gets an opportunity to turn to synchronized fragments of sounding speech and its spelling.

Synchronization of sound files is carried out employing ELAN software. The project team is grateful to prof. Dobrushina N.R., Head of International Laboratory of Language Convergence, for arranging a training session on ELAN software usage.

***Metamarkup***

 Metamarkup of the corpus is based on the surveyed responses to two questionnaires: a sociolinguistic one, developed at the Institute of Linguistics of the Russian Academy of Sciences, based on the questionnaires by O. A. Kazakevich and used to study languages of minor peoples in the Russian Federation [Sociolinguistic questionnaire, 2018], and a bilingual language questionnaire, based on the questionnaire of language experience and language proficiency developed by Marian V., K. Blumenfeld H., Kaushanskaya M. [Marian et al.,2007].

The sociolinguistic questionnaire includes 41 questions with more detailed information about the author of the text: time and place of birth, residence, education, professional activity, information about relatives by different types of kinship, the nature of the acquisition and use of languages he/she speaks.

The bilingual language questionnaire includes 14 blocks of parameterization of the nature and type of a bilingual’s command of interacting languages: languages ​​are ranked according to the extent of their use, the order of languages acquisition, the amount of time the subject uses languages “at present time” (around the period he/she is interviewed), the choice of languages ​​when reading and communicating with another person. The subjects also provide language use details – the time of each language acquisition, the time spent in the language environment, the self-assessment of the language level and factors that stimulated acquiring each of the languages ​​spoken by the bilingual, the preferred topics and areas of communication for speaking in each of the languages.

Thus, metamarkup is structured based on the informant’s Russian language proficiency level, which determines its status: whether the Russian language, according to the self-assessment of the informant, is native (mother) or non-native; used in the period of the interview in different forms of communication (active vs. passive), the order of language acquisition (L1 vs. L2); the sphere of preferential use (written vs. oral; household vs. official vs. aesthetic vs. others). The same parameters are used to assess other language(s) the bilingual (polylingual) employs.

Analysis of the respondents’ responses to the questionnaires helps to compile a generalized social portrait of a bilingual and distinguish between age groups that differ in their ethnic language proficiency and L1 / L2 ratio. The first age group consists of the bilinguals aged 18 to 35, the second age group – the bilinguals aged 36 to 65, the third age group – the bilinguals aged over 65. A comparison of the data obtained makes it possible to trace the interaction dynamics of the languages spoken by the minor peoples in Southern Siberia and the Russian language as the macro-intermediary language in the region, and to identify existing trends in the further development of the process.

The authors believe the text discoursive parameters mapping should be as detailed as possible, therefore the metamarkup also includes text characteristics of 1) discourse type (institutional, personal), 2) communication type (monologue, dialogue, polylogue), 3) genres (interview, discussion, conversation, report, report, story), 4) topics. The information about the text also includes the data on the time and place of the interview, text size in words.

The first, most general parameter of text qualification is its belonging to one of discourse types accepted in in world and Russian linguistics – institutional or personal. The recording as a whole can be characterized by switching the discoursive register. Thus, fragments of an interview that record an interviewer, a student, or a university teacher, keeping to strict regulations of the questionnaire, and an informant’s corresponding performance, are interpreted as institutional, tagged [DInst]. However, the interview can turn into a friendly conversation, recollections. Such texts extracts are classified in the corpus as related to personality discourses and tagged [DPers].

Further, in the corpus, an opposition of communication types is marked: monologue [CTMlg], dialogue [CTDlg], or polylogue [CTPlg]. Alternating dialogues and monologues dominate in the corpus.

The ways of recording oral speech for a database predetermine the predominance of speech genres [GInt], conversation [GConv], recollection [GRec], etc. in the corpus.

The metadata is completed with an indication of the topic of the text fragment, currently identified as typical are the following topics:

* stories about one’s life, recollecting memorable periods, e.g. childhood, school years (tagged [TPerd]);
* stories about memorable events, e.g. wedding, child bearing, divorce, journeys (tagged [TEvt]);
* stories about people, e.g. friends, neighbours, relatives (tagged [TPeople]);
* stories about environment, e.g. nature, weather, describing village, town, living conditions (tagged [TEnv]);
* stories about ethnic culture, folklore, e.g. feasts, traditions, shamans (tagged [TEC]);
* stories about daily routine, e.g. hunting, fishing, working in the house (tagged [TDR]).

Morphological annotation of the text is carried out using Mystem, the console program of the Yandex company, morphological information is limited by the ability of this program.

***Error annotation***

We developed a system for error annotation based on the existing marking system in the Russian Learner Corpus [RLC, 2018], as well as in the Contact-Related Russian Speech Corpus, created as part of the project "Dynamics of language contacts in a circumpolar region", studying the speech of bilingual native speakers of minor languages in the North of Siberia and the Far East [Corpus, 2017]. We use the term “speech standard deviation” (SSD), since the corpus contains the recordings of speakers of different forms of the Russian national language: literary, dialectal colloquial, whose speech practices correspond to the standards of the corresponding subsystems, not being “errors". As a basic point, the standards of Russian written literary speech were taken, on the basis of which the principles of grammatical writing of the language were formed in the Grammatical Dictionary of the Russian Language by А.А. Zaliznyak, which is the basis of Mystem automatic morphological marking. The corpus is a collection of oral speech recordings, which results in a number of deviations from the written speech standard, determined by the oral, spontaneous nature of communication. Fixing such deviations, in our opinion, expands the informational possibilities of the corpus by annotating the markers of conversational communication.

The basic principle of annotation is level representation of SSDs. For the time being a system of one- or two-degree qualification of the deviation type is adopted; in a small number of cases, the third degree of specification is presented, which is motivated by the need to reflect the transfer phenomena:

* the first degree – marking language level where the deviation occurs ([Phon] – phonetics, [Morph] – morphology, [Synt] – syntax, [Lex] – vocabulary, [Disc] – discourse);
* the second degree – marking a particular linguistic phenomenon within the language level, to which the deviation is related e.g. [PhonAcc] – phonetics, accent; [MorphInfl] – morphology, inflexion; [SyntGov] – syntax, government, etc.;
* the third degree – marking a particular kind of the deviation, e.g. [SyntAgrGen] – speech standard deviation in agreement by gender, where [Synt] refers to the first degree (syntactical language level), [Agr] refers to the second degree (agreement - a way of expressing syntactical relations within a phrase), and [Gen] refers to the third degree (deviation in the agreement according to the grammatical gender).

The second important characteristic of the system of marking speech standard deviations is the marking the source of deviation: intralinguistic, interlingual influence. The bilinguals whose speech practices are represented in the corpus belong to different age and social groups, have different education and speak of different variants of the Russian language: literary, dialect, dialect-colloquial, colloquial, tagged [Reg] (regional).

The conclusions about the internal system source of deviations from the standard of the literary language are made on the basis of the data recorded in the works on the description of the subsystems of the Russian language under consideration [Araeva 1991, Russian colloquial speech. 1983, Russian dialects of the Middle Ob valley Vol.1,1984, Vol. 2 1989, Zemskaya et al. 1981].

Conclusions about the external source of influence are formed on the basis of data on the peculiarities of the structures of the Turkic languages, published data on transfer in this area of interlingual interaction [Comparative historical grammar of Turkic languages. Phonetics 1984, Comparative historical grammar of Turkic languages. Morphology 1988, Comparative historical grammar of Turkic languages. Vocabulary 2001, Gordeeva 1962, Gordeeva 1965]. Deviation from the speech standard due to the influence of the native language is tagged [Trsf].

SSD annotation cases:

(1) *И дорога эта / котора* [MorphGov; MorphAff; Reg] (которую) *сделали* *“And the road that was made”*

(2) *Много, гыт, шорска блюда* [MorphGov; Int; MorphAff; Reg] (шорских блюд) *готовили “A lot of Shor dishes are said to have been prepared”*

(3) *Сделала тут такой яма* [SyntGov; Int; SyntAgrGen; Int] (такую яму) *“She made here such a hole”*

(4) *Какой мне позвоночник править, я сама-то не трогаю, боюся [*MorphAff; Reg] *(боюсь) “I can't fit my backbone, I don’t touch it myself, I’m scared”.*

A marked extract from the text: [DInst-Pers], [Dlg], [GInt], [T People]

(5) *Нет, бабушка обряд [Lex] (наряд) никогда ..Как [SyntCon] у нее было платье с такими каемочками, а чтобы так [SyntLessW]. Она, во-первых, не курила, не пила, поэтому, наверно, с бабушками не обращалась [Lex] (не общалась), потому что в поселке бабушки пили и с трубкой курили, [MorphGovPrep] и всегда ругались, вот помню, и дрались между собой. Напьются, О..!.,{смех} бабушка говорит: зачем, гыт [Disk], пить, грит[Disk]? Она не пила никогда. Почти сто лет прожила. Она девяносто девять [WLess] (девяносто девять лет) с половиной прожила. И в своем уме. А мы щас [Phon] (сейчас) чё [Phon] (что) положим, уже забыли. “* *No, the grandmother rite [Lex] (outfit) never .. As a [SyntCon] she had a dress with such laces, and so it was [SyntLessW]. First, she didn’t smoke, didn’t drink, so, probably, she didn’t contact [Lex] (didn’t communicate) because elderly women drank and smoked in the village, [MorphGovPrep] and always cursed, I remember, and fought among themselves. Get drunk, Oh ..!., {Laughter} grandmother says: why, gyt [Disk], drink, grit [Disk]? She never drank. Almost a hundred years lived. She was ninety-nine [WLoss] (ninety-nine years old) and a half lived. And was in her right mind. And we right now [Phon] (now) what [Phon] (what) we put, have already forgotten.”*

 ***Corpus database structure***

The final structure scheme of the database consists of 9 entities (Fig.1).

Fig. 1.Corpus database structure

The performance of the scheme was tested on the local server, metamarkup based on the DBMS PostgreSQL was verified (Fig. 2).



 Fig. 2 Search in accordance with the database markup on the local server

To test the automatic markup, the texts of respondents were taken from all three age groups of representatives of Russian-Shor, Russian-Khakass and Russian-Tatar bilinguals. Russian-Tatar bilinguals: the first group - 3 hours 17 minutes of sounding, 22,500 words, the 2nd group - 2 hours 25 minutes, the text volume - 18,500 words, the third group - 1 hour 12 minutes, about 9,000 words; Russian-Shor bilinguals - the first group: the duration of the recording is 46 minutes, the volume of the text in words is about 6000; the second group - the duration of the recording is 2 hours and 14 minutes, the volume of the text is about 17,600 words; the third group: the duration of the recording is 2 hours, the volume of text in words is about 13,300; Russian-Khakass bilinguals; the first group is 3 hours 9 minutes, the volume of text in words is about 22,500; in the second group, the recording duration is 2 hours 5 minutes, the text volume in words is about 21,000. In the third group, the recording duration is about 2 hours, the text volume – about 26,700 words.

Marking (metamarkup, morphological annotation, error annotation) was done with ELAN. The work was carried out on the integration of the sound file and the text relative to the timeline, installed into the database presented in Fig. 3.

Fig. 3. An example of annotating a text produced by a Russian-Tatar bilingual in ELAN

Such integration is possible employing ELAN intended to form tier-like system and to annotate texts synchronized along time scale. Figure 3 shows 9 tiers:

1) «Text itself» (text\_pure) – tapescript of the informant’s recording, divided into гееукфтсуы;

2) «Words» (text\_words). Word splitting in a sentence that is relative to the timeline and sentence;

3) «Morphological annotation» (words\_morph) – word annotation based on Mystem;

4) «Speech standard deviation» (error) – annotating SSD revealed in the process of communication.

5) «Corrections» (word\_correct). Correction of the informants’ SSDs. This tier is necessary to provide correct operation of the database search algorithm. For example, the informant produced a speech deviation while pronouncing the word “*sosedna*”. Without correction, in searching for the word “*sosednyaya*” (*neighbouring*) the deviation will be missed. So, corrections are considered to be necessary to optimize the search for SSDs and the normalization of texts.

The tiers 6 – 9 annotate text passages as belonging to a particular

6) discourse type ([DInst-Pers]),

7) communication type ([СТDlg]),

8) genre ([GInt]), and devoted to a

9) particular topic ([T People]).

Thus, the corpus data represent the features of the regional version of the Russian language existing in the interlanguage contacting area. The system of annotation and meta-markup allows exploring SSD types, the social and psycholinguistic factors’ affecting their actualization. The corpus can be a source of research for a regional version of spontaneous oral speech.

***References***

1. Araeva L.A. (1991), Kuzbass dialects in their current state, [Govory` Kuzbassa v ix sovremennom sostoyanii], Coordination meeting of the departments of the Russian language of universities in Siberia, the Urals and the Far East on the problems of studying Siberian dialects [Koordinacionnoe soveshhanie po problemam izucheniya sibirskix govorov kafedr russkogo yazy`ka vuzov Sibiri, Urala i Dal`nego Vostoka], Krasnoyarsk
2. Comparative historical grammar of Turkic languages. Phonetics (1984), [Sravnitel`no-istoricheskaya grammatika tyurkskix yazy`kov. Phonetika], М
3. Comparative historical grammar of Turkic languages. Morphology (1988), [Sravnitel`no-istoricheskaya grammatika tyurkskix yazy`kov. Morphologiya], М
4. Comparative historical grammar of Turkic languages. Vocabulary (2001), [Sravnitel`no-istoricheskaya grammatika tyurkskix yazy`kov. Leksika], М
5. Corpus of contact-related Russian speech (2017) [Korpus kontaktno-obuslovlennoj russkoj rechi] available at <http://web-corpora.net/tsakorpus_russian_nonst/>
6. Gordeeva O.I. (1962), Some observations on the Russian language acquisition by the Tatars under the conditions of the old-time Russian environment [Nekotory`e nablyudeniya nad usvoeniem russkogo yazy`ka tatarami v usloviyax russkogo starozhil`cheskogo okruzheniya], Tomsk State University Bulletin [Ucheny`e zapiski Tomskogo gosudarstvennogo universiteta], Vol. 40, pp. 75-80
7. Gordeeva O.I. (1965), Some regularities of the influence of the native language on the acquired language in the process of the formation of bilingualism (based on Russian and Tatar Siberian dialects) [Nekotory`e zakonomernosti vliyaniya rodnogo yazy`ka na usvaivaemy`j yazy`k v processe stanovleniya dvuyazy`chiya (na materiale russkogo i tatarskogo sibirskix govorov], Tomsk
8. Marian V., Blumenfeld H.K., Kaushanskaya M., (2007), Language Experience and Proficiency Questionnaire (LEAP-Q), Speech Language and Hearing Research, Vol. 50 (4), pp. 940- 967, available at: http://www.bilingualism.northwestern.edu/leapq/ (date of access: 07.09.2018).
9. Rezanova Z.I., Nekrasova E.D., Miklashevsky А.А., (2018), Investigation of psycho-linguistic and cognitive aspects of language contacting in the project “Linguistic and ethnocultural diversity of Southern Siberia in synchrony and diahrony: interaction of languages and cultures” [Issledovanie psiholingvisticheskih i kognitivnyh aspektov yazykovogo kontaktirovaniya v proekte «Yazykovoe i etnokul'turnoe raznoobrazie Yuzhnoj Sibiri v sinhronii i diahronii: vzaimodejstvie yazykov i kul'tur»] Rusin, № 2 (52), pp 107 – 117 DOI: 10.17223/18572685/52/8
10. RLC: Russian Learner Corpus (2018), available at http://www.web-corpora.net/RLC[/](http://www.web-corpora.net/RLC/)
11. Russian colloquial speech. Phonetics. Morphology. Vocabulary. Gesture (1983) [Russkaya razgovornaya rech`. Fonetika. Morfologiya. Leksika. Zhest] Nauka, M
12. Russian dialects of the Middle Ob valley (1984) [Russkie govory` srednego Priob`ya] V.1, ed. V.V. Palagina, Tomsk
13. Russian dialects of the Middle Ob valley (1989) [Russkie govory` srednego Priob`ya] V.1, ed. V.V. Palagina, Tomsk
14. Sociolinguistic questionnaire, available at URL: http://ilingran.ru/main/departments/ural-altaic (date of access: 07.09.2018).
15. Zemskaya E. A., Kitajgorodskaya M. V., Shiryaev E. N. (1981), Russian colloquial speech. General questions. Word formation. Syntax. [Russkaya razgovornaya rech`. Obshhie voprosy`. Slovoobrazovanie. Sintaksis], Nauka, M
1. The research was supported by the Government of the Russian Federation (№ 14.Y26.31.0014). [↑](#footnote-ref-1)