Sociolinguistic Variability of Russian Everyday Speech: A Corpus-Based Study

Natalia Bogdanova-Beglarian[1], Olga Blinova[1,2], Tatiana Sherstinova[2,1], Ekaterina Baeva[1],
Daria Gorbunova[1], Tatiana Popova[1]

[1]Saint Petersburg State University,
[2]National Research University Higher School of Economics,
Saint-Petersburg, Russia
{n.bogdanova, o.blinova, t.sherstinova, e.baeva}@spbu, {dgorbunova2, tipopova13}@gmail.com}

Abstract—**The paper presents recent results of a multilevel analysis of representative corpus data, conducted in order to identify key speech parameters (lexical, morphological and syntactic) that can diagnose some social/biological characteristics of a speaker or, more broadly, a modern Russian urban sociolect. The study is based on the everyday Russian speech corpus "One Speaker's Day". Specific data were obtained on the analysis of the annotated subcorpus of 289,205 tokens, which includes recorded "speech days" of 57 men and 48 women, which were the research participants, as well as speech fragments of 87 men and 139 women, which were their interlocutors. Thus, the total number of speakers in the subsample amounts to 144 men and 187 women. The article also begs the question of Data Mining approach usability to the subcorpus and possibilities of further research using machine learning. The results obtained are important for the optimization of speech technologies systems, for theoretical understanding of linguistic processes, as well as for monitoring various social processes taking place in modern Russian metropolis.**

I. Introduction

It has long been acknowledged that language cannot be studied in isolation from the speaker, e.g.: "Since language only exists in human society, then, in addition to the mental side, we must always take into account the social side" [1: 348]. Any person's speech reflects characteristics of the society while performing a certain role assigned to him: "any individual is a prisoner of his own language; outside his class, he reveals himself with every spoken word, each word reveals his full self and flaunts it along with his entire history. Thanks to his language, a man is open to deciphering; he is betrayed by the very truthfulness of the linguistic form, which never wants to lie about itself" [2: 347]. Sociolinguistists have been traditionally observing the "correlations, in which certain social parameters, stratification or situational, are independent variables, whereas linguistic phenomena depend on them. These correlates could show either complete or incomplete functional dependence" [3: 481]. The aim of this research is to reveal such dependences.

The relationship between social class and speech has been studied by many scientists (G. N. Putnam and E. M. O'hern, K. R. Scherer, H Giles, N. Coupland, S. W. Gregory, J. J. Bradac and R. Wisegarver).

However, there seems to be little consensus in what exactly can be defined as class or status [4]. A. Hollingshead uses a two-factor model to define the speaker's status, the two factors being the education and occupation of the head of the household [5]. The famous language researcher W. Labov (1966) assigns to each of his respondents a special socioeconomic index, calculated from the results of a sociological survey which took into account the person's years of education, the occupation of the family breadwinner, and family income. From that data W. Labov drew up four social classes and studied the speech of these groups [6].

In Russian linguistics a speaker's social status is traditionally defined as "the relative position of an individual or group, determined by its social characteristics (economic status, profession, qualification, education, etc.), natural characteristics (gender, age, etc.), as well as prestige and rank in power structure" [7: 789].

The concept of social status in Russian linguistic studies includes various characteristics of the speaker, both biological and social: 1) socially significant differences between people are of a biological and social nature and are typified in the system of social characteristics of an individual, 2) social characteristics of an individual form a hierarchy in accordance with the values of a particular community of a certain period, 3) -social characteristics of the individual are heterogeneous in various respects, are grouped into characteristic complexes and can be measured, 4) the social status of a person is revealed in the role, distance and normative characteristics of behavior due to socio-economic and cultural-ethical factors of public life, 5) role, distance and normative characteristics of behavior are reflected in strategies and means of non-verbal and verbal communication "[8: 322]. The indicators of social status have been claimed to be present even in short speech acts [9]. Information about the speaker's age, sex, state of health, mental attitude and native geographical region is often conveyed with the communicative message and can be revealed even in some 15 seconds of speech [10, 11]. Given that, the purpose of this research was to study the everyday (everyday and professional) speech of large social groups in a modern Russian city aspiring to find out how, for example, the speech of a worker differs from the speech of an engineer or teacher, is there any difference between the speech of office workers and that of intelligentsia, between the language of the youth and the elderly, etc. Are there any specific features that are able to diagnose the speaker's status, thus immediately revealing this status to the listener? Was Mikhail Bakhtin right when saying that "All words smell like a profession, a genre,

<...> a certain person, generation, age, day and hour" [12: 106]?

In other words, the result of the undertaken research should be a set of social and biological features of the speakers with a particular sociolect. This can be essential for a variety of purposes: from purely scientific tasks of developing systems for automatic processing of sound signals, speech recognition and synthesis, to forensic linguistics, for example, in order to combat terrorism, when a potential intruder could be identified by "sociological portrait", compiled by an expert based on a speech sample.

II. Problem statement

A multidimensional analysis of extensive speech subcorpus has been carried out to test a number of linguistic parameters for their ability to diagnose various sociolects of Russian society. With some of these parameters being rather traditional in sociolinguistic research (gender, age, level of education, profession) (see the works of T. I. Erofeeva, E. V. Erofeeva, M. M. Bakhtin, N. B. Vakhtin, E. V. Golovko, V. I. Belikova, L. P. Krysin, E. I. Goroshko, A. V. Kirilina, N. and J. Coupland, H. Giles, J. Fishman, E. B Ryan, D. Tannen, P. Trudgill, etc.), others were introduced for Russian speech analysis for the first time, i. e. the speaker's level of speech competence [13], his social status and place of birth and the place of the longest residence in Russia. Moreover, it is the first time in Russian speech research that all these factors are taken into account simultaneously. The research is unique in analyzing the Russian language in its most natural form (everyday dialogues, not limited by laboratory conditions or speech tasks), in all possible communication conditions, including both everyday and formal speech. The material for analysis was the corpus of Russian everyday speech "One Speaker's Day" (ORD) [14, 15, 16, 17]. In creating the corpus a 24-hour continuous recording was used for all speech production of research participants and their interlocutors. It is worth mentioning that ORD is the first Russian corpus assembled this way. A similar method of speech recording has been used in Japanese linguistic studies [18, 19], and in the preparation of materials for the oral subcorpus of British National Corpus [21]. At the moment, the corpus comprises more than 1 million words in text transcripts and includes more than 1,400 hours of sound recordings of domestic and institutional communication of representatives of various Russian urban social groups [22].

The speech recording material from a significant sample of speakers was the basis for the analysis of urban "multilingualism" on different linguistic levels, taking into account various communicative situations. This contributes not only to understanding linguistic processes, but might be useful in observing important urban social trends.

III. Research Data

During the study, expert manual annotation was carried out at the lexical and discursive level, in order to highlight style-specific, professional, slang vocabulary and neologisms. At the morphological level, automatic morphological annotations were carried out, followed by manual data correction (disambiguation); as well as the identification of rare and pragmatically marked forms (for example, vocative case) including annotation of agrammatic, occasional forms. At the syntactic level, expert annotation of linear structures (word order) for nominal groups and verbal groups has been performed, the number of left and right dependencies for verb groups has been determined, and specific syntactic phenomena of oral speech (parcellation, ellipsis, disruptors, self-corrections) have been identified. As a result, the annotated size of the corpus amounted to 289,205 tokens, which include "speech days" of 57 men and 48 women (informants), and 87 men and 139 women (communicants). The total number of speakers in the subsample is 144 men and 187 women.

4 stages of pilot annotation of the material were carried out. The first pilot annotation was based on a sample, which included 16,000 word usages from the ORD corpus and was carried out in parallel by 5 experts, according to the rules developed at the preparatory stage. During annotation, an expanded list of PM functions was used, while annotators selected the main one from them and put the corresponding tag in first place in the "PM Function" level box. At the "PM Comment" level, some additional features of markers usage were noted.The tags themselves represented the designation of the corresponding function. Possible new PMs, as well as various options already available in the list, were noted with the help of a special mark at the level of "PM Comment". An analysis of the results of the first pilot annotation showed that the markup instruction needs some updates. In the course of preparing the instructions for the second pilot annotation, it was decided to use a shorter list of PM functions and the list of the main and additional functions in the same alphabetical order, since almost every PM in multilingual speech turned out to be multifunctional, and the hierarchy of the functions performed by it during markup was not always built up uniquely. Analysis of the results made it possible to optimize the methodology and develop more effective instructions for markers. The revised methodology was successfully tested at the second stage of annotation and remained unchanged at the third and fourth stages. The third pilot annotation was carried out on the SAT sub-corpus (15,000 word usage). It was carried out for a preliminary assessment of the characteristics of the use of PM in monologic speech. The fourth pilot annotation was based on ORD sub-corpus (60,000 word usage). It was carried out for a preliminary assessment of the characteristics of the use of PM in dialogical speech and allowed some conclusions to be drawn about the features of the use of PM in male and female speech. At the end of each phase of the pilot annotation, expert proofreading of pragmatic markup was carried out, the list of allocated PM was revised and supplemented. At the moment, the working list of PM variants includes 450 units, which are variants of 53 basic structural types.

As the next step of sociolinguistic analysis of annotated material, the quantitative results have been drawn for numerous social groups. It has been organized into the following clusters, at lexical, morphological and syntactic levels:

(1) three groups by level of education (higher, incomplete higher, secondary special),

(2) three groups by level of speech competence (high, medium, low),

(3) five groups by place of birth / main residence,

(4) two gender groups,

(5) three age groups (junior, middle, senior),

(6) ten professional groups (blue-collar workers, engineers, military personnel, academics of natural sciences and liberal arts, educators, service providers, IT specialists, office employees, creative intelligentsia),

(7) five status groups.

Consequently, a list of features has been formed that distinguish speech of one social group from another at all levels.

A. Lexical level

The lexical analysis of the corpus material was carried out on 77,240 word tokens. The IPM (items per million) of different lexical groups was calculated for the corpus as a whole and separately for all the identified sociolects. This is the list of labels used to mark the data in annotation:

(1) chronological labels: OLD – archaic, NEO – neologisms;

(2) phraseological labels: IDIOM – set expressions;

(3) functional labels: SPESH – special vocabulary,

(4) stylistic (including emotional and expressive) labels: OFST – formal style, NOF – informal style (colloquial vocabulary), SRV – stylistically reduced vocabulary, EUPH – euphemisms, BRAN – swear words;

(5) pragmatic labels: ETI – etiquette formulas (greetings, farewells, apologies);

(6) word-building labels: DIM – diminutive, NA – nomina agentis (name of the agent).

Below is a fragment of a table containing the results of the lexical annotation of the data (see Table I).

TABLE I. FRAGMENT OF THE LEXICAL ANNOTATION TABLE

Sfile	Phrase	Scode	lexmarks: ST	lexmarks: FORM	lexmarks: FUNC
ordS1225	as'ka / as'ka zarabotala / on skazal chto sajt ne rabotaet // eto ja vyjasn'u //	M1-S12	NOF		SPESH
ordS1225	ja eto iznachal'no kak by // *V na stadii iznachal'nogo / *P vot prosto kogda ona *N ne sokhranilis' / no my s Oksankoj% tak zhe dolgo / *P (e...e) udivl'alis' / *P s chem eto sv'azano / s che... s chego vdrug Tan'a% tak vot / a Ol'a% nachala peredo mnoj pr'amo begat' // ona (...) v mae kogda ja prinesla zajavlenie / ona tak bojalas' chto ja ujdu / *V i / potomu chto ona bojalas' / chto ja Oksanku% s soboj voz'mu / *P bojalas' chto / kak by ona ostanets'a voobshche odna / kto zhe budet rabotat' ! *P zdes' // *P to est' a potom vot / khochet / pust' sidit kak by //	S12	NOF		
ordS1225	no ja () pon'ala to est' / nastol'ko to est' / *P nu on... / *P vot (i...i) / ja by vot esli chestno govor'a / ja (...) ushla by uzhe iz etoj firmy / *P kogda vot men'a vot / tak vot () pinali // *P prosto pinali //	S12	NOF		
ordS1225	prosto udivl'ajus' kak by jejo (...) kolossal'nomu terpeniju / *P i ponimaju vot / pochemu kak by vot (...) chto ona dejstvitel'no kak by delaet / i tak dalee // a kogda dejsi... / dejstvitel'no my ne znaem / kto chto delaet / eto bol'shaja beda / eto bol'shaja beda / eto (...) tol'ko v tom chto rukovodstvo / (e...e) net u nas meroprijatij / *V my vs... drug protiv druga // my tam schitajem chuzhie zarplaty / my tam jeshch'o chto-to // *V vs'o tak skopishche takoe / komok / ot etogo vot / *P azh toshno // mne dazhe protivno / vot pravda / mne ne () mne vot ochen' //	S12	OFST		
ordS1225	no u nikh / (e...e) on skazal () Andrej% / chto u nikh / bol'she set' chem nasha // i (...) BiSiSHuz$(?) kak by / *V i bol'she formata / i (e...e) *V to est' nu () p... pon'atno / to chastichno sp... Sportmaster$ / i vot drugie marki // kakie marki / vot ja () tozhe / *P navernoe / budu obshchat's'a / *V nu *V (e...e) ja tak otvykla ot sobesedovanij / bojus' vot *S ne ponravit's'a //	S12	OFST		

The most common top four groups in the corpus were as follows (hereinafter, unless otherwise noted, the numbers in parentheses indicate IPM): SPESH (12 597), ETI (4 195), NEO (285) and OLD (207). This indicates the prevalence of special vocabulary in everyday speech, including both work and everyday conversations. A similar ranking distribution of these 4 groups in all sociolects turned out rather similar, with sporadic but very significant deviations.

For example, there have been no neologisms (NEO) in the speech of senior citizens and students, whereas the speech of a number of sociolects didn`t include archaic words at all (OLD); these were businessmen, unemployed pensioners, representatives of natural specialties, residents of the South Federal District (SFD) and Uzbekistan (UZ). In the speech of a number of groups, neither neologisms nor archaic vocabulary were found: these are children (who in this research acted only as communicants), as well as groups of engineers, IT specialists, educators, workers with a low level of speech competence and secondary special education, as well as residents of most regions of Russia (except for the above SFD and UZ).

In the speech of the liberal arts academics, with a complete set of top types of vocabulary, archaic words prevailed over neologisms; in the speech of office workers and those with incomplete higher education there were neither neologisms nor archaisms, but the top recorded group was the etiquette formulas. Finally, from the top vocabulary groups only special words appeared in the speech of representatives of power structures.

B. Morphological Level

In order to annotate different parts of speech and perform lemmatization, we used morphological analyzer MyStem [24] with a standard set of 13 basic tags: V – verb, S – noun, SPRO – pronoun/noun, NUM – numeral, A – adjective, APRO – pronoun adjective, ANUM – ordinal numeral, ADV – adverb, ADVPRO – pronoun adverb, CONJ – conjunction, INTJ – interjection, PART – particle, PR – preposition.

The annotation was performed using the latest version of Mystem available on GitHub (Version 3.1 for Windows). The following commands were selected for annotation:

• -n ("print each word on a new line"),
• -l ("do not print the original word forms, only lemmas and grammemes"),
• -i ("print grammatical information"),
• -g ("combine wordform information with one lemma"),
• -d ("apply contextual removal of homonymy").

The list of constant and inflectional grammatical features of word forms is also provided by the analyzer [24].

Top-5 of the frequency lists of parts of speech in the corpus material looks approximately the same in all social groups:

1) verbs (V; in the general frequency list 16.7 %),
2) nouns (S; 15.5 %),
3) pronouns (SPRO; 14.4 %),
4) particles (PART; 14.3 %);
5) conjunctions (CONJ; 7.7 %).

Deviations from this distribution are rather rare yet significant addressing to our research goals.

Thus, particles (PART) are slightly more frequent (compared to SPRO) in the speech of middle-aged informants (Age=2), as well as in the speech of men and the retired people. They are especially frequent in the speech of businessmen (rank 1, ahead of nouns and verbs), as well as in the speech of residents of the Southern Federal District (also rank 2, but after SPRO; the verbs here "moved" to 3rd place).

In a number of sociolects, in the top 5 of the frequency list of parts of speech, in contrast to the general trend, there are adverbs (ADV) (men, representatives of law enforcement and creative professions, as well as informants with incomplete higher education and residents of the South and North-West Federal Districts (SFD and NWFD), as well as prepositions (PR) (engineers, workers, informants with secondary specialized education and residents of Moscow and the Siberian Federal District – SFD).

Verb (V) is the most frequent group in most sociolects, with rare exceptions. In the speech of businessmen and residents of the Southern Federal District, it is ranked 3; in the speech of engineers, workers, representatives of creative professions, speakers with secondary specialized education, speakers of a high level of speech competence and residents of the Northwestern Federal District – rank 2. Most often in these situations the verb gives way to a noun.

Below is a fragment of a table containing the results of morphological annotation (see Table II).

TABLE II. FRAGMENT OF THE MORPHOLOGICAL ANNOTATION TABLE

SFILE	WORD	SCODE	POS	GRAM
ORDS60-05	*G	S60		
ORDS60-05	TY	S60	SPRO	NOM, 2D SG
ORDS60-05	PRISHOL	S60	V	SG, M, PF, ACT, PAST
ORDS60-05	?	S60		
ORDS60-05	TEBE	S60	SPRO	DAT, 2 SG
ORDS60-05	KASHU	S60	S	

C. Syntactic Level

The syntactic analysis of the corpus material was carried out in three directions: identification and statistics of the syntactic features of everyday speech, statistics of syntactic structures and characteristics of the verb positions.

The most typical syntactic features in almost all sociolects of Russian society were as follows: speech disruptors (CUT) (2.73 % of the total number of identified structures), ellipsis (EL) (1.69 %), parcellation (PARC) (1.33 %) and speaker self-correction (COR) (0.48 %). In the speech of gender groups and most other specified social groups, the distribution was in full accordance with the data on the corpus as a whole. However, some sought-after exceptions were discovered, claiming the status of potential diagnostic sociolinguistic parameter.

Thus, EL turned out to be more frequent than other structures in the speech of office and creative workers, students, speakers with incomplete higher education and residents of the Siberian Federal District. In the speech of workers and speakers with a high URC, parcellation tops the list, "overtaking" both the disruptors and the ellipsis, and in the speech of the citizens of the Northwestern Federal District it was self-correction. In the speech of businessmen and residents of Moscow, at the syntactic level no other specific features except for disruptors have been found at all. Finally, in a number of sociolects (middle and senior age groups, liberal arts, natural scientists, engineers, law enforcement agents and retired pensioners), the top 4 syntactic features included non-projective (inverted) constructions (*sovetskuju chital / znajesh' literaturu; tvoj u men'a nomer*) – as a rule, with a rank of 4, "squeezing out" parcellations or self-correction. It is noteworthy that disruptors are present in all sociolects, even in the speech of children, which was not specifically studied, but fell into the attention zone as the speech of communicants.

The top 4 syntactic structures in the whole corpus were indivisible sentences (consent formulas) (Y; *aga, da, ladno* and the like) (2.7 %), single verb (V) (2.5 %), noun phrase (NP=S) (1.9 %) and the standard SV predicative structure (pronoun / noun-subject + verb-predicate) (1.6 %). Only in the speech of women and speakers from the senior age group do pure verb structures take the first position, in their top-4 being also pragmatic markers (D) like *vot, nu, eto*, etc., that help the speaker build the sentence.

Finally, an analysis of the verb positions, i.e., the nature of syntactic structures, showed the exact same picture for all sociolects. The most typical situations (top 5) were the following: 0V0 (a single verb, without any additional components – *Pojdu*), 1V0 (one unit on the left of the verb – *Ja poshol*), 2V0 (two units on the left of the verb *On ne prishol*), 1V1 (one unit on both the left and right of the verb – *Eto bylo vchera*) and 0V1 (one unit on the right of the verb – *Pojdu domoj*). It can be observed that left branching in everyday spoken speech significantly exceeds the right one, unlike in written language, where symmetrical structures tend to dominate [25].

Below is a fragment of a table containing the results of syntactic annotation of the data (see Table III).

TABLE III. FRAGMENT OF A TABLE CONTAINING SYNTACTIC ANNOTATION RESULTS

i posidet' polepit' / zdes' jego pokaraulit' //	B (NP=x S) O3 HV	vrachi	S	2	0	20	
etot ? sejchas posmo... //	B B			0	0	00	
i poetomu ja bolee-menee vot na segodn'a osvobodils'a / chtoby segodn'a zanyat's'a tam () zavtra etim remontom //	Conj2 B S B V			3	0	30	
*nu prosto vot jeshch'o smotrite kak by jest' vot takaja model' stul'chikov // jesli vas pugaet vot tkan' / *P vot smotrite kak //*	Conj2 S V			1	0	10	PARC
*i kogda / *P ulichali / bylo strashno nelovko // *P za rechevuju ili pis'mennuju //*	Conj3 S Pred(=A)			0	0	00	
nu on mne sk... () predlozhil pr'am paren' govorit v et... magazine / govorit...	M(NP=S) Intr V {B}	Gul`%	S	0	1	01	CUT
smotri //	S D V			1	0	10	

IV. CONCLUSION

The paper presents the results of multidimensional linguistic annotation of the corpus material (the corpus of Russian everyday speech) and a statistical description of the specifics of Russian urban sociolects, at the lexical, morphological and syntactic levels. In the course of this large-scale research, we strive to probe the previously suggested [22] statistical hypotheses on the significance of the differences between the analyzed sociolects.

A number of scientific methods was used to form a subcorpus and information system, to annonate and process speech data, with later proceeding to identify significant diagnostic features of various social groups at different language levels:

- methods of speech technologies (a set of methods for processing and analyzing speech processes; building speech databases);
- methods of corpus linguistics and information methods for building multimedia databases;
- methods of linguistic annotation;
- methods of discursive analysis;
- methods of quantitative linguistics, statistical methods of data processing and testing of statistical hypotheses.

To work with textual corpus annotations, we used the multi-level linguistic annotation by ELAN [26], which allows associating direct decoding with a sound wave. In this study, version 5.4 with advanced functionality was used: it supports performing glossing directly inside ELAN and compare annotations of the same files made by different annotators, etc.

Text annotation of the corpus has been done manually; however, working with annotation files * .eaf, a number of original discoveries was applied. In particular, to correct technical errors of files before their automatic processing, we designed and used the Corrector software utility, which is able to generate an error log reflecting inconsistencies in the transcribing speech overlapping phenomena at the Phrase and Speaker level, to locate empty boxes with missing information at the Speaker level, etc.

In addition, the Eafer utility (another application developed by our research group) was used to automatically perform the "dilution by speaking" operation and receive annotation files with as many Phrase levels as there are speakers.

As a result of the research, the following sociolinguistic parameters can be defined as potentially diagnostic.

It turned out, for example, that liberal arts academics have a tendency to use many archaic words, while office workers and speakers with incomplete higher education lean towards etiquette formulas.

In the speech of businessmen, middle-aged speakers, as well as men and the retired, particles were more frequent than in other sociolects.

Ellipsis is typical for the speech of office and creative workers, students and speakers with incomplete higher education. Parcellation prevails in the speech of workers and speakers with high URC. In the speech of speaking middle and senior age groups, liberal arts academics, natural scientists, engineers, law enforcement agents and the retired, a predominance of inverted structures is clearly observed, being more frequent than parcellation and self-correction.

Again, this data can be important for a variety of purposes: from purely scientific tasks of developing systems for the automatic processing of sound signals, speech recognition and synthesis, to forensic linguistics and linguistic didactics.

VIII. FURTHER RESEARCH

Despite the fact that the results of the analysis of the corpus material in all considered aspects (vocabulary, morphology, syntax), obtained at the first stage of the study (2016 [22]) and in this work, confirm each other quite well, the final set of diagnostic signs for the mentioned so far can only be considered potentially diagnostic, requiring another confirmation on the expanded corpus material. This remains the prospect of studying everyday Russian spoken speech in the chosen direction. Traditional linguistic techniques are not sufficient for the analysis of real time spontaneous speech even in transcripts. Therefore most research in the area considers data mining tools from the linguistic domain as main for mining big amounts of data as corpora [20].The first and central question to the approach if data-mining methods are able to generate and then verify the existing research results. The second one is the ability to lead the linguist to further linguistically interesting patterns emerging from the data of non-standard speech corpora. One more crucial issue in this endeavor is how to discover those linguistic features that are good indicators of sociolinguistic differences, provided they exist. As you can see above, we have explored a set of features potentially distinguishing between the different social classes, on the one hand, but they must be verified, on the other hand. In our future work we plan to carry out more analyses on the basis of aggregated linguistic information using the described approach in order to explore more concrete parameters of variation.

ACKNOWLEDGMENT

The presented research was supported by the Russian Foundation for Basic Research, project 17-29-09175 “Diagnostic Features of Sociolinguistic Variation in Everyday Spoken Russian (based on the Material of Sound Corpus)”.

REFERENCES

[1] I. A. Baudouin de Courtenay, Selected Works on General Linguistics, Moscow: USSR Academy of Sciences, 1963, Vol. 1, pp. 348-350.

[2] R. Barthes, Writing Degree Zero, Semiotics [Collection of articles. Translations] / Yu. S. Stepanov (ed). M.: Raduga, 1983, pp. 306-349.

[3] A. D. Schweitzer, Sociolinguistics // Linguistic Encyclopedic Dictionary / V. N. Yartseva (ed.), Moscow: Soviet encyclopedia, 1990, p. 481.

[4] S. Ash, Social class. In The Handbook of Language Variation and Change, J. K. Chambers, P. Trudgill and N. Schilling-Estes (eds.). Malden, Mass.: Blackwell, 2002, pp. 402-422.

[5] A. B. Hollingshead, Two Factor Index of Status Position. New Haven: Yale University Press, 1957, pp. 1-11.

[6] W. Labov, The Social Stratification of New York City. Washington, DC: Center for Applied Linguistics, 1966.

[7] Sociology: Encyclopedia / Comp. A. A. Gritsanov, V. L. Abushenko, G. M. Evelkin, G. N. Sokolova, O. V. Tereshchenko. Moscow: Knizhny Dom, 2003.

[8] V. I. Karasik, Language of Social Status. Moscow: IYa AS of SSSR, Volgograd Pedagogical Institute, 1991.

[9] G. N. Putnam and E. M. O'hern, The Status Significance of an Isolated Urban Dialect. Language, 31(4). 1955, pp. 5-32.

[10] L. S. Harms, Listener Judgments of Status Cues in Speech. Quarterly Journal of Speech, 47(2), 1961, pp.164-168;

[11] D. James, Listener judgments of status cues in speech: A replication and extension, Speech Monographs, 39:2. 1972, pp. 144-147.

[12] M. M Bakhtin. Questions of Literature and Aesthetics. Moscow: Khudozhestvennaya literature Publ, 1975.

[13] Speech Corpus as a Base for Analysis of Russian Speech. Collective Monograph. Part 1. Reading. Retelling. Description [Zvukovoj korpus kak material dl’a analiza russkoj rechi: kollektivnaja monografia. Chast’ 1. Chtenie. Pereskaz. Opisanie]. N. V. Bogdanova-Beglarian (ed.). St. Petersburg, 2013.

[14] A. Asinovsky, N. Bogdanova, M. Rusakova, A. Ryko, S. Stepanova, T. Sherstinova, The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation, in: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI, vol. 5729. Springer, Berlin-Heidelberg, 2009, pp. 250-257.

[15] N. V. Bogdanova-Beglarian, T. Yu. Sherstinova, O. V. Blinova, G. Yu. Martynenko, An Exploratory Study on Sociolinguistic Variation of Spoken Russian. SPECOM 2016. Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, 2016, pp. 100-107.

[16] N. V. Bogdanova-Beglarian, T. Yu. Sherstinova, O. V. Blinova, O. B. Ermolova, E. M. Baeva, G. Ya. Martynenko, A. I. Ryko, Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech, in: Ronzhin, A. et al. (eds.) SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, 2016, pp. 659-666.

[17] N. V. Bogdanova-Beglarian, O. V. Blinova, T. Yu. Sherstinova, G. Ya. Martynenko, Corpus of Russian Everyday Speech "One Day of Speech": present state and prospects, in: Proceedings of the V. V. Vinogradov Institute of Russian language. Vol. 21. Russian National Corpus: Research and Development / A. M. Moldovan, V. A. Plungyan (eds.). Moscow, 2019, pp. 101-110.

[18] T. Sibata, Study of Language Existence within 24 Hours] // Linguistics in Japan. Moscow: Raduga, 1983, pp. 134-141.

[19] N. Campbell, Speech & Expression; the Value of a Longitudinal Corpus, LREC 2004. Lisbon, 2004, pp. 183–186.

[20] Han and M. Kamber, Data Mining: Concepts and Techniques. 2006.

[21] Reference Guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services. L. Burnard (ed.), 2007. [Web: http://www.natcorp.ox.ac.uk/docs/URG/2007].

[22] Everyday Russian Language: Functioning Features in Different Social Groups. Collective Monograph. N. V. Bogdanova-Beglarian (ed.)., St. Petersburg, 2016.

[23] Mystem Web: https://tech.yandex.ru/mystem/

[24] G. Ya. Martynenko, The Syntax of a Live Spontaneous Speech: the Symmetry of Linear Orders // Corpus Linguistics. 2015. Proceedings of the International Conference, in V. P. Zakharov, O. A. Mitrofanova, M. V. Khokhlova (eds.). Saint Petersburg: SPBU, 2015. pp. 371-378.

[25] ELAN Web: https://tla.mpi.nl/tools/tla-tools/elan/.