Design and Data Collection for the Accentological Corpus of the Russian Language

E. Grishina*, S. Savchuk*, A. Poljakov**

*Institute of the Russian Language RAS, Moscow, Russia

**Scientific Technical Centre "Informregistr", Moscow, Russia

E-mail: rudi2007@yandex.ru, savsvetlana@gmail.com, pollex@mail.ru

Abstract

Accentological corpus provides a researcher an opportunity to study word stress and stress variation, which are very important for the Russian language. Moreover, Accentological corpus allows studying the history of the Russian language stress development.

The research presents the main characteristics of Accentological corpus available at ruscorpora.ru. Corpora size, type and sources of text material, the way it is represented in the corpora, types of linguistic annotation, corpora composition and ways of their effective use according to their purposes are described.

There are two zones in the Accentological corpus. 1) The zone of *prose* includes oral texts and films transcripts, in which stressed syllables are marked according to the real pronunciation. 2) The zone of *poetry* contains texts with marked accented syllables, so it is possible to define the exact word stress using special rules.

The Accentological corpus has four types of annotations (metatextual, morphological, semantic and sociological) and also has its own accentological mark-up. Due to accentological annotation each word is supplied with stress marks, so a user can make queries and retrieve the stressed or unstressed word forms in combination with grammatical and semantic features.

1. Introduction

The interest to the spoken language data is constantly increasing in corpus linguistics, as referred to (Rayson & Mariani, 2009). There are common recommendations and standards concerning preparation and representation of oral texts in the corpus (TEI, EAGLES). Meanwhile each national corpus (British, Czech, Slovak, Russian, American etc.) offers its own way of texts selection and corpus architecture.

As far as the Russian National Corpus is concerned, we have refused the idea of creating a "universal" spoken corpus suitable for all kinds of oral speech study - from phonetic aspects to discourse analyses. That's why there is not one but three spoken sub-corpora within the RNC, with their own features and spheres of application. These are a) Spoken sub-corpus (Grishina, 2005a; Grishina, 2009a; Grishina & Savchuk, 2009), b) Accentological sub-corpus (Grishina, 2009c) and c) Multimedia Corpus (is under development) (Grishina, 2005b; Grishina, 2009b).

The paper presents the main parameters of the Accentological corpus.

2. Accentological corpus of Russian

Accentological corpus gives a researcher an opportunity to study word stress. This information is very important for languages with non-fixed stress. The Russian language is one of them. It has a very complicated stress system. Russian stress has the following features: firstly, it is non-fixed, which means any syllable may be stressed (for example, *zo'loto 'gold'*, *voro'na 'raven'*, *boroda' 'beard'*); secondly, it is mobile, which means that it may shift from one part of the lexeme to another as a result of inflexion or word formation (e.g., *zo'loto*, noun *'gold'* – *zoloto'j*, adj *'gold'*, *zoloti't'*, verb *'cover with gold'*, *pozolo'ta 'gilding'*; *ruka'* 'hand' (n, f, nom, sg) – *ruki'* (n, f, gen, sg), *ru'ku* (n, f, acc, sg), *ru'ki* (n, f, nom, pl).

Moreover, the Russian stress system is in the process of rearrangement, and significant accentological changes take place practically under our eyes. Change of accentological system is the main factor of occurrence of stress variation in forms of inflection or derivation. It also causes changing of stylistic evaluation: which variants can be accepted as standard and which are not. That is why normative recommendations given in certain reference books and dictionaries may differ greatly from each other and from actual usage (Mustajoki, 1990). This brings extra difficulties in learning Russian, especially for those who learn it outside Russia.

The researchers using sociolinguistic methods in studying stress variation, for example, in the form of surveys of native speakers come to a conclusion that there are no satisfactory descriptions of Russian stress system because all of them are based on lexicographical sources and reflect recommended usage. So they are far away from what 'people are actually saying' (Lagerberg, 2007; Marklund Sharapova, 2000; Ukiah, 2002).

Accentological corpus, large and representative, allows obtaining information of word stress not from dictionaries, but from real texts. From the very beginning the Accentological corpus was planned as a kind of

diachronic corpora, which would allow studying the history of the Russian stress. Meanwhile, there are no limitations to the size of the corpus (except for technical capabilities). The goals of the corpus define its design and the criteria for data collection.

2.1 Criteria for text selection

There are two zones in the Accentological corpus.

1) The zone of *prose* includes the oral texts and the films transcripts, in which stressed syllables are marked according to the real pronunciation. The main criterion for including a text in the corpus is the availability of a corresponding record (the quality of a record ought to give us possibility to verify the transcript). We are interested both in the accentological standards of the literary language and in their variants which emerge in course of time. The text annotation makes it possible to characterize any variant from the point of view of a sphere of functioning, a genre and a speaker and to evaluate its frequency and regularity. Thus, the prose zone contains some examples of spontaneous everyday speech, public colloquial speech of different levels of spontaneity (TV and radio speech, political speeches, academic spoken speech, sermons, *etc.*), movie and radio plays transcripts, reading aloud. The earliest records of this zone date from the beginning and the first decades of the 20th century (gramophone records of L.N. Tolstoy's letters, political leaders' speeches, records of speeches made at the First congress of writers in 1934, and movies of the 1930s). In perspective some of the accented written texts (e.g., books and manuscripts of the 18th and 19th centuries, and later – even older texts) may be included in this zone.

2) The zone of *poetry* contains texts with marked accented syllables, so one can define the exact word stress using special rules. Specially annotated poetic texts of the 18th-20th centuries are included in this zone and still continue to be added. At present this zone mainly reflects the history of the Russian stress, as the corpus contains poetry written before the 20th century.

2.2 Structure and composition of the corpus

Accentological corpus cannot be called balanced in the usual sense of the term, which is used to describe a large corpus. A representative and balanced corpus suggests that it includes a certain proportion of texts belonging to various aspects of language functioning. Accentology is not mainly interested in language use in general, but in contextual use of a certain set of lexemes that form the unstable, moving part of the accentual system. Therefore, a balance of an accentological corpus should be expressed in the fact that it contains speech referring to the different functional areas (public, non-public, professional), varied in terms of gender, age, education level of the speakers, as well as regionally and chronologically.

Nowadays the Accentological corpus contains more than 8.7 million tokens. Texts distribution among the two zones and according to time periods is listed below in Tables 1 and 2.

Zone		**Tokens**	**Percentage**
Poetry		4082253	46.8
Prose	Movie speech	4298000	49.3
	Public speech	290290	3.4
	Private speech	17877	0.2
	Reading aloud	25277	0.3

Table 1: Texts proportion in the zones of Accentological corpus

Zone	**Poetry**	**Prose**
1700-1799	926720	
1800-1899	2933190	
1900-1949	222343	492150
1950-1979	-	2031752
1980-1999	-	1024708
2000-2008	-	1017900

Table 2: Distribution of texts according to the date of creation

2.3 Types of annotation

The Accentological corpus is supplied with four types of annotation which are used in the RNC and also has its own accentological mark-up.

Metatextual annotation marks a text as a whole and includes information regarding author's name, sex, age or date of birth, date of text recording / creating etc. Also the parameters that are specific to each zone (prose and poetic) are used. These are *genre, meter, clause, rhyme, strophe type* for the poetic zone and *text type* for the prosaic one.

Morphological information is assigned to a word-form and consists of four groups of tags: 1) lexeme (a dictionary form of the lexeme and the part of speech to which it belongs); 2) a variety of the lexeme's grammatical features, known as word-classifying features; 3) a variety of the word-form's grammatical features, known as word-altering features; 4) information concerning non-standard forms of the word-form, orthographic variations, *etc*.

Sociological annotation is specific to the spoken corpora only. It is assigned to different speaker's utterances and characterizes a word usage from the point of view of sex and age of a speaker (if this information is available). Sociological annotation allows a user to create his/her own sub-corpora by various parameters or their combinations: by a speaker's sex (so a user could create a sub-corpus of feminine or masculine spoken language); by a speaker's age (for example, a user can create a sub-corpus of teenagers' phrases); by a speaker's year of birth (this option is available only for movie transcripts, so you could select the phrases by the actors born in 19th century); by an actor's name (for example, you can create a sub-corpus of Eugene Leonov's phrases).

Apparently, sociological annotation may be supplemented with metatextual annotation which makes it possible to select texts by one speaker and include his/her name and year of birth in the description of the text. It is clear, that if a) there are more than one speaker, b) speakers cannot be named because of ethical reasons, c)

their age is unknown or speakers are of very different age, this information cannot be included in the description of the text. In this case all we can refer to is sociological annotation.

Due to *accentological* annotation each word is supplied with stress marks, so a user can make different kinds of queries and retrieve the stressed or unstressed word-forms in combination with grammatical and semantic features.

2.4 Language data preparation and presentation method

Preparation of texts for the Accentological corpus is performed in several steps.

The first step includes decoding of the audio files, an orthographical normalization and editing of the transcripts. Then we accentuate the transcripts in manual mode, using Accentuator software (by A. Poljakov) which operates with the data of the embedded lexicon. This lexicon includes the database of normative Russian dictionaries, but it is also amplified by the corpus developers. At the third stage an expert listens to the audio records and corrects the transcripts. As a result we get a text which reflects the real pronunciation.

(1) [Maya, Zhanna Kerimtajeva, fem, 35, 1953] Ón govorít / poká ne otremontírujete trubú vo dvoré / ón ne vkl'úchit. *[Yu. Mamin, V. Vardunas. Fontan, film (1988)]*.

As we have mentioned above, in the poetry zone the same annotation as in the Poetry corpus is used. The special Metrics program (by A. Poljakov) marks up strong beats (potentially stressed syllables) in a poem. As a result we get a text which looks as follows:

(2) On idEt v vorotA, on uzhE na kryl'cE, on vzoshEl po krutYm ***stupen'Am*** na plosh'Adku i vIdit: s pechAl'ju v licE odinOko-unYlaja tAm *[V.A.Zhukovskij (1822)]*

(3) Vot nAsh gerOj pod'jEhal k sEn'am; shvejcAra mImo On strelOj vzletEl po mrAmornYm ***stupEn'am***, rasprAvil vOlosA rukOj, voshEl. *[A.S.Pushkin (1823-1824)]*.

In these two citations the stresses of the form dat. pl. of the noun *stupen'am* 'steps' are different, which shows the coexistence of these variants in the beginning of the 19th century.

In example (2) up beats (ictuses) are more frequent than word stresses. In this case we should exclude all *impossible* stresses (which are not presented in any dictionaries or reference books) and take into consideration *possible* stresses only. For the word-form *mra'mornym* the only possible stress is the stress on the first syllable, for the word-form *volosa'* (n, m, nom, pl) – the stress is on the ending, but the stress on the first syllable would characterize this word-form as gen. sg.

2.5 Estimated usage and prospects for development

The Accentological corpus is one of the specialized corpora in the framework of the RNC. It is intended for the researches in a specific sphere of the Russian accentology. A rather small size of the corpus is quite sufficient to let a researcher study accentological trends, to verify the hypotheses. However, usage of the Corpus transcends the sphere of accentology, as the Russian accentuation relates to the Russian morphology and semantics. Thus the corpus data would be useful for researchers of morphology, phonetics and prosody, syntax and semantics of Russian (Grishina, 2009c).

Another important sphere of application is the lexicography and the codification of literary language. Normative recommendations, including those concerning position of stress, are usually based on the data extracted from dictionaries and researchers' linguistic experience, whereas the corpus gives us the possibility to observe stress patterns in real texts during a long period of time, to test and correct recommendations.

Furthermore, the corpus can be useful in language teaching and learning. Russian stress is difficult to study, especially when learning Russian as the second language (Andrews, 2001; Kerek, 2009). Thus, corpus data can be used as a reference material and as a material for compiling exercises.

The following examples illustrate standard tasks that can be tackled using accentological corpus material.

Example 1. In modern Russian the word *kamen'* (*stone)* has a wavering declination paradigm. The basic version has a movable stress – scheme 2*e, according to (Zaliznjak, 1977): in the singular and in nom./accus. cases pl. the stress is on the stem, while in the oblique cases pl. the stress is on the inflexional ending. Another option is the constant stress on the stem in both numbers, i.e., scheme 2*a. It is interesting to see how this system has established itself over the past three centuries. With work on accentological corpus yet to be completed, it is only natural that there are chronological gaps remaining therein; for example, insufficient or lacking data for the early 20th century and hardly any information available from the poetic texts for the 20th century. However, the material now available is sufficient to understand the general trends. Figure 1 shows the relationship between the accentological corpus options.

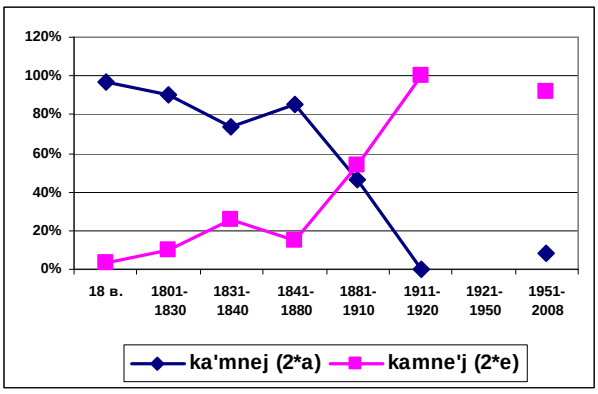

Figure 1: The correlation between stress variants of gen. pl. forms of the word *kamen'* in different time periods

As you can see, the early 21st century situation is mirror-opposite to that of the 18th – beginning of the 19th century: the new scheme (2*e: *ka'mni, kamne'y)* became completely predominant, while the previous scheme (2*a:

ka'mni, ka'mney) prevailed absolutely in the 18 century. As seen from the diagram, the turning point was the period from 1881 to 1910 when the accent patterns were almost equifrequent. The remarkable leap in the use of the pattern 2*e in 1831-1840 is connected with Mikhail Lermontov who was inclined, as analysis of the system of accents in his poetry shows, to use "progressist" accentological models.

Example 2. It is well known that the word *muzyka* (*music*) changed its accent (it was shifted from the second to the first syllable) during the 19th century. It is interesting to find out how it occurred. Because this word was of relatively high frequency in the Russian poetry of the 18th – 20th century, the data we obtain from accentological corpus are representative enough to show that the accent shifting process was a gradual one in this period and the turning point was the first third of the 19th century.

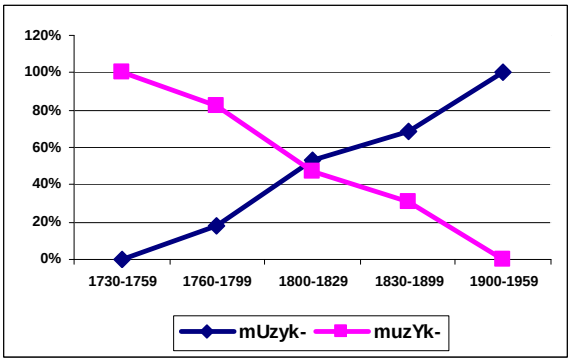

Figure 2: The correlation between stress variants of the word *muzyka* (*music*) in different time periods

The process was smooth and gradual as it embraced equally all the cases, as shown in Table 3.

	Case form ratio	mUzyk-	muzYk-
Total	100	60	40
Nom	38	60	40
Gen	24	43	57
Dat	3	100	0
Acc	12	64	36
Ins	18	69	31
Prep	5	80	20

Table 3: Distribution of stress variants in case forms of the word *muzyka* (*music*)

As can be seen from Table 3, significant deviations from the mean values (60% - accent on the first syllable and 40% on the second) are only due to the low-frequency cases of this word (dative and prepositional). Interestingly, the genitive case proved the most "conservative" in accepting the new accent. It should be noted that, as far as this parameter is concerned, I.A.Krylov's creative work can be regarded as "revolutionary" enough: Krylov never used the stress on the second syllable and therefore was far ahead of his time.

(4) NevEzhda v fIzikE, a v ***mUzykE*** znatOk, uslYshal sOlovjA, pojUschegO na vEtke, i khOchets'A jemU imEt' takOgo v klEtke. [I.A. Krylov. Pavlin I solovej (1788)].

(5) …Khoz'Ain ***mUzykU*** l'ubIl i zAmanIl k sebE sosEda pEvchikh slUshat'. [I.A. Krylov. Muzykanty (1807)]

A more sudden shift (which occurred at the boundary between the second and third thirds of the 19th century) exhibits a change in accent in the singular masculine of the short form of the adjective *sil'nyi* (*strong) (si'len* vs. *siljo'n)*, as it is seen in figure 3.

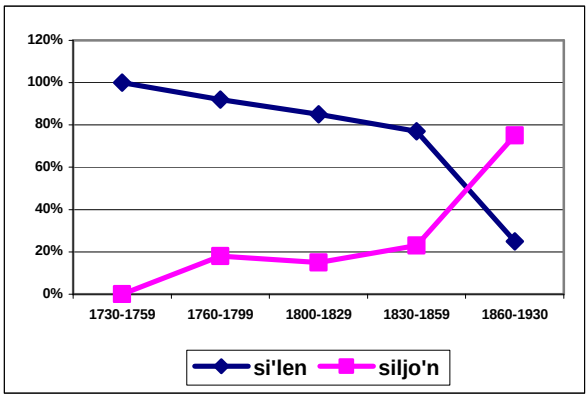

Figure 3: The correlation between stress variants of the short form of the adjective *strong* in different time periods

It should be noted again that in Ivan Krylov's work the replacement of *si'len (strong)* by *sil'on* occurred much earlier versus other poets: all examples of *sil'on* with the accented second syllable in 1760-1829 can only be encountered in Krylov's texts. (Of course, you can find therein the *si'len* option too.)

In the near future the expansion of the Corpus size and the increase of text variety are expected. According to this plan poems of the 1st half of the 20th century will be included to the zone of poetry. The zone of prose will be replenished with the texts belonging to different spheres of spoken communication and created in various time periods. We can mention academic lectures, interviews and TV talk-shows, sports comments, sermons, political speeches, narratives, private conversation etc.

3. Acknowledgements

This work was supported by: Fundamental Research Programs of Branch of History and Philology RAS "Genesis and coordination of social, cultural and language communities"; "Text in sociocultural environment: levels of historical, literary and linguistic interpretation"; Presidium RAS Fundamental Research Program (project "The Russian language of the 18th century: corpus-based study of lexical and morphological variability"); RFBR (grant 08-06-00371-a).

References

Andrews, Edna (2001). The Russian Reference Grammar. Available at: http://www.seelrc.org:8080/grammar/

mainframe.jsp?nLanguageID=6

Grishina, E.A. (2005a) Ustnaja rech v Natsional'nom korpuse russkogo jazyka. In *Natsionalnyj korpus pusskogo jazyka: 2003–2005*. Rezul'taty i perspektivy. Moscow. Available at: http://docs.google.com/View?id=df52fjjj_12fcbpqbdq

Grishina, E.A. (2005b) Dva novych projekta dl'a Natsional'nogo korpusa russkogo jazyka: multimedijnyj podkorpus i podkorpus nazvanij. In *Natsional'nyj korpus pusskogo jazyka: 2003–2005. Rezul'taty i perspektivy*. Moscow. Available at: http://docs.google.com/View?id=df52fjjj_13fhtjh5ct

Grishina, E.A. (2009a) Natsional'nyj korpus russkogo jazyka kak istochnik svedenij ob ustnoj rechi. *Rechevyje tekhnologii, 3*. Available at: http://docs.google.com/View?id=df52fjjj_34g9d9w2dg

Grishina, E.A. (2009b) Multimedijnyj russk'ij korpus: problemy annotatsii. In *Natsional'nyj korpus pusskogo jazyka: 2006-2008. Novyje rezul'taty i perspektivy*. SPb: Nestor-Istorija. Available at: http://docs.google.com/View?id= df52fjjj_363wxt76dk

Grishina, E.A. (2009c) Korpus "Istorija russkogo udarenija" In *Natsional'nyj korpus pusskogo jazyka: 2006-2008. Novyje rezul'taty i perspektivy*. SPb: Nestor-Istorija. Available at: http://docs.google.com/View?id=df52fjjj_37ghmg36cb

Grishina, E.A., S.O. Savchuk (2009) Korpus ustnyh tekstov NKRJa: sostav i struktura. In *Natsional'nyj korpus pusskogo jazyka: 2006-2008. Novyje rezul'taty i perspektivy.* SPb: Nestor-Istorija. Available at: http://docs.google.com/View?id=df52fjjj_39gh8wsffv

Kerek, E., P . Niemi. Russian orthography and learning to read. In *Reading in a Foreign Language*, April 2009, Vol. 21, No. 1, pp. 1–21.

Lagerberg, R. (2007). Variation and Frequency in Russian Word Stress. *ASEES, Vol. 21*, Nos. 1-2 (2007), pp. 165-176

Marklund Shaparova, E. (2000). *Implicit and Explicit Norm in Contemporary Russian Verbal Stress*. Uppsala: Uppsala University (Acta Universitatis Upsaliensis. Studia Slavica Upsaliensia).

Mustajoki, A. (1990). Unifitsirujuts'a li chislovyje podparadigmy udarenija russkich suschestvitel'nych na *–a*? In *Wiener Slawistischer Almanach*, 25/26, pp. 311-326.

Rayson, P., J. Mariani. (2009) Visualizing corpus linguistics. In *Corpus Linguistics 2009*. 20-23 July 2009. Abstracts, p. 201. Liverpool.

Ukiah, N. (2002). The stress of Russian nouns in *–a* and *–я* of Zaliznjak's pattern f (*губа'* type). *Australian Slavonic and East European Studies*, 16/1-2, p. 1--39.

Zaliznyak, A. (1977/2003) Grammaticheskij slovar' russkogo jazyka. M: Russkije slovari.