D.O.I. http://dx.doi.org/10.1590/2176-4573e55822

ARTICLES

Principles of the St. Petersburg Phonological School in Speech *Corpora* Design / *Os princípios da Escola Fonológica de São Petersburgo para a elaboração de* corpora *de fala*

Pavel Skrelin[*]
Tatiana Kachkovskaia[**]
Daniil Kocharov[***]
Vera Evdokimova[****]
Uliana Kochetkova[*****]

ABSTRACT

The paper discusses the main principles in designing and annotating speech *corpora* within the framework of the Saint Petersburg phonological school, and provides examples of using *corpus* data in phonetic research. One of the major principles that we follow is to analyse the speech material at all levels: from segmental to intonational, including speech disfluencies. During segmental phonetic annotation, we suggest listening to each speech sound in isolation (without knowing its context) and relying on spectrographic data. At the syllabic tier, it is crucial to reflect resyllabification. During prosodic annotation, we suggest to rely on listener's perception of the intonation pattern first, then analyse the actual melodic curves. A speech *corpus* with multi-level annotation that follows these principles is a valuable source of phonetic data — as segmental and prosodic factors are in constant interaction with each other, and one cannot analyse units of one annotation tier without reference to other tiers.

KEYWORDS: Phonetics; Phonology; Speech *corpus*; Speech annotation; St. Petersburg phonological school

RESUMO

O artigo discute os princípios fundamentais de elaboração do projeto e anotação de corpora *de fala no âmbito da Escola Fonológica de São Petersburgo e fornece os exemplos de utilização de dados de vários* corpora *na pesquisa em fonética. Um dos princípios fundamentais é analisar as amostras em todos os níveis: desde o segmento até a entoação, incluindo as disfluências da fala. Durante a anotação fonética, sugerimos ouvir cada som isoladamente e confiar nos dados do espectrograma. Na anotação silábica, é crucial considerar a ressilabificação. Durante a anotação prosódica, sugerimos confiar na percepção do ouvinte e analisar as curvas melódicas. Um* corpus

[*] Saint Petersburg State University, Department of Phonetics, Saint Petersburg, Russia; https://orcid.org/0000-0002-8355-7378; skrelin@phonetics.pu.ru

[**] The manuscript was prepared while the author worked at Saint Petersburg State University, Department of Phonetics, Saint Petersburg, Russia; https://orcid.org/0000-0002-8588-9165; kachkovskaia@phonetics.pu.ru

[***] The manuscript was prepared while the author worked at Saint Petersburg State University, Department of Phonetics, Saint Petersburg, Russia; https://orcid.org/0000-0002-7858-5331; kocharov@phonetics.pu.ru

[****] Saint Petersburg State University, Department of Phonetics, Saint Petersburg, Russia; https://orcid.org/0000-0001-9742-5299; postmaster@phonetics.pu.ru

[*****] Saint Petersburg State University, Department of Phonetics, Saint Petersburg, Russia; https://orcid.org/0000-0003-1792-6064; u.kochetkova@spbu.ru

de fala que segue esses princípios é uma fonte valiosa de dados fonéticos, uma vez que *os fatores segmentais e prosódicos estão em constante interação e não se pode analisar as unidades de um nível de anotação sem fazer referência aos outros.*

PALAVRAS-CHAVE: Fonética; Fonologia; Corpus *de fala; Anotação fonética; Escola Fonológica de São Petersburgo*

Introduction

Nowadays, much of the phonetic research is based on *corpus* data (Liberman, 2019). The first large speech *corpora* contained detailed manual annotation. For example, TIMIT, which includes read sentences recorded from 630 English speakers from various parts of the USA, is carefully segmented into speech sounds (Garofolo et al., 1993). Another well-known corpus, BURNC, is notable due to its prosodic annotation using the ToBI system (Ostendorf et al., 1995).

With time, as *corpora* grew in size, it became obvious that manual annotation was too time-consuming. The development of speech processing tools, however, enabled the researchers to apply automatic annotation and segmentation procedures based on ASR. With extremely large *corpora*, such as Librispeech (Panayotov et al., 2015) and VoxPopuli (Wang et el., 2021), manual annotation became practically impossible. VoxPopuli contains 400,000 hours of unlabeled European Parliament speech data in 23 languages, of which only 1,800 hours are transcribed. Despite the absence of labelling, such databases are suitable for phonetic research (see, for examples, Chodroff; Wilson, 2017).

This recent trend, however, by no means decreases the importance of annotated speech *corpora* for the phonetic science. Resources like TIMIT and BURNC are still the gold standard for research of segmental and suprasegmental phenomena in English.

Most of speech *corpora* are provided with annotation manuals which typically describe terminology and notation. The annotation principles are discussed casually and raise a lot of additional questions in the minds of thoughtful readers. This paper is aimed to discuss the principles of speech *corpora* design and annotation that are applied at the Department of Phonetics, St. Petersburg State University.

The paper begins with a brief historical outline describing the first speech collections stored at the department. Then follows a detailed discussion of the principles of segmentation and annotation at each particular tier — from phonetic to intonational and paralinguistic. Finally, we provide brief descriptions for the most notable speech *corpora* developed at the department; this section ends with examples of phonetic research performed based on these *corpora*.

1 Historical Background

The Department ("Cabinet") of Experimental Phonetics in Saint Petersburg was founded in 1899. In over a hundred years, phonetic research here has used the same instruments as other similar laboratories in the world: from tuning forks and kymographs to large speech *corpora* and models of speech production. The present-day Department of Phonetics and Methods for Teaching Foreign Languages,[1] St. Petersburg State University, has a long history of processing large collections of speech recordings. Between the 1950s and the 1980s, phonetic research relied on sound archives containing detailed descriptions of the recordings, and kilometres of film with oscillograms and spectrograms. The largest sound archive designed in the 1970–80s included speech recordings produced by speakers of various regional variants of Russian. The recordings were made in over 70 major cities of the Soviet Union, and at each location at least 20 male speakers participated — mainly, students residing in the given region or the republic who spoke the regional (dialectal) variant of Russian or the local variant bearing some traits of their native language. The illustrative material characterising the phonetics of each variant of Russian was gathered into a separate audio tape, and the results of this phonetic research were published in the late 1980s (Bondarko; Verbitskaya, 1987).

At the end of the 1980s, the development of the Mashinnyj fond russkogo jazyka [*Russian Computer Fund*] started. A substantial part of it was intended to be the phonetic data bank representing a digital micro-model for the sound system of Russian (Bondarko et al., 1992). The progress in developing the *Foneticheskij fond russkogo jazyka* [Russian Phonetic Collection] and the results of related research were regularly published in the special journal *B'ulleten' foneticheskogo fonda russkogo jazyka* [The Bulletin of the Russian Phonetic Collection]. The journal came out since 1988 in Bohum, Germany, and later partly in Saint Petersburg. There were two editors: Prof. Christian Sappok from

[1] Further on, we will use the short variant: "Department of Phonetics."

Ruhr-Universität Bochum, and Prof. Lia V. Bondarko, the Head of the Department of Phonetics in Leningrad State University (later, St. Petersburg State University), Russia. The development of the *Russian Computer Collection* was interrupted in the early 90s. However, by that time the creation of *Russian Phonetic Collection* had been almost finished, and in 1993 the research group published Appendix 3 for *the Bulletin of the Russian Phonetic Fund* with the title *A Collections of Sound Units of Russian Speech* (Appendix 3, 1993).

Appendices for the Bulletin had always included an audio tape with the research material, that was later replaced by a CD. *Appendix 3 for the Bulletin of the Russian Phonetic Fund* was published with a paper by Bondarko (1993) on the sound system of Russian, along with a detailed and illustrated description of all the components (modules) of the data bank. Module *The Syllable* included spectrograms and recordings of all the 186 Russian CV and V syllables, with segmentation into sounds. Module *The Word* contained 150 words with non-transparent orthography, 250 words from the Basic Learner's Dictionary (Paperno; Leed, 1988), and a frequency dictionary of Russian. Module *The Text* included a phonetically representative text, a two-page text containing the most frequent Russian phonemes and syllables, and a related dialogue illustrating Russian intonation system. All the audio materials, except for the frequency dictionary, were recorded from four speakers (2 males, 2 females) representing two main variants of the standard Russian pronunciation: those of Moscow and Leningrad (St. Petersburg).

2 Principles for Design and Annotation of Speech *Corpora*

The ideas that emerged during the work on *the Russian Phonetic Fund* and the experience obtained in those years enabled to form the framework for designing phonetic data banks for other languages spoken in Russia. This task required re-evaluation of the principles for data collection, data organization, quality assessment via auditory experiments or technical equipment, and software and hardware selection. The main principles underlying the *Russian Phonetic Collection* were adjusted according to the specific features of languages to be described. In the end, the language data bank acquired the following structure:

1. audio data;

2. phonetic features of each of the minimal meaningful units of the language (phonemes/syllables);

3. phonetic structures of word forms;

4. automatic grapheme-to-phoneme converter;

5. phonetic features of intonational units.

An annotated speech *corpus* requires annotation of speech data, including segmentation and labeling of segmental and suprasegmental speech units. Further on, we will discuss the principles for annotating and labelling speech *corpora*.

Layering Principle. In speech *corpora* annotation, we follow the principle of strict layering: each unit of a smaller tier must be in*corpora*ted fully into one and only unit of a higher tier. As a result, we do not allow higher boundaries to lie inside a phoneme. This principle enables easier automatic processing, although requires a number of extra segmentation rules.

In connected speech, we often observe phoneme omissions, insertions, or fusions, which require special attention. In case of omissions, the absent phoneme often leaves something behind: e.g., a labialized vowel, when not pronounced, still causes rounding of the preceding consonants (e.g., Russian "существование," existence: [sʷʃʲːistvʌˈvanʲii̯]).[2] With insertions, e.g. vowel insertions in some consonantal clusters, we have to decide what counts as a phoneme, as insertions are often very short (e.g., Russian "корабль," ship: [kʌˈrabᵊlʲ]).[3] Fusions are by definition hard to divide. When two identical sounds occur one after another, there is no clear boundary between them; we may place the phonemic boundary right in the middle of the sound, but this is more disputable when we deal with geminated stops as plosion in such cases is produced only once (e.g., in Russian "оттуда," therefrom: [ʌˈt̚_tudʌ]). These issues can be partly resolved by annotating speech at different segmental tiers. This allows to describe one phoneme as consisting of several sounds, and a single sound corresponding to several phonemes.

[2] Interestingly, when a vowel is omitted, the perceived number of syllables does not change: in this example the sequence [sʷ] still forms a syllable on its own. Thus, rhythmical structure of the word remains intact.

[3] Typically, such vowel insertions do not change the number of perceived syllables. That is, similar to the previous example, rhythmical structure of the word remains intact.

An obvious and only exception from the layering principle is observed at the syllabic tier due to the phenomenon of resyllabification (e.g., Russian “брат Ани,” Anya’s brother: [ˈbrat ˈanʲi], [bra-ta-nʲi]). For some languages, division into rhythmical feet may also defy this principle (e.g., English “come again”: [ˈkʌmə-ˈgen]).

Setting the Boundaries. Segmental boundaries must be detected as precisely as possible. This guarantees high precision of boundaries at higher annotation tiers. There are a number of published recommendations for segmentation, e.g. (Turk et al., 2012). The segmentation principles may be based on the particular tasks for which the *corpus* was constructed. In general, the labels are placed at boundaries of the physical realisations of allophones. Segmentation should meet the following criterion: the resulting allophones can be successfully transplanted into other words containing the same type of sound (Skrelin, 1999).

Two Tiers for Phonetic Annotation. The “acoustic” phonetic tier is the commonly accepted phonetic tier which can be found in many of the well-known speech *corpora*. It is produced by listening to the analysed sounds in isolated context and by using instrumental methods (spectrograms). The principles behind this tier are: (1) de-lexicalization and (2) taking into account the acoustic properties of sounds, to ensure the maximal precision and objectivity of the transcription.

De-lexicalzation is aimed to solve problems with phonetic interpretations caused by good phonological ear of the listener (phonetician). Without this principle, the annotator’s decision is likely to be influenced by their knowledge of the phonemic content of the pronounced word. (Bondarko et al., 1974). The annotator’s decisions must be based on precise data on the physical properties of sounds. This is especially crucial for vowels, where we should rely on formant values (for data on Russian vowels, see Evdokimova et al., 2020).

“Perceptual” phonetic transcription is produced by listening to the word as a whole. Such transcription enables to reveal the perceived particularities of the speaker’s pronunciation, including those specific to a certain region. This type of transcription differs from the acoustic phonetic transcription because it is based on the annotator’s knowledge of the pronunciation standard (or the main dialect). As a result, this tier contains only perceptually relevant information that is easily heard (e.g., Russian “водяной,” water spirit: [vʌdʲa̱ˈnoi̯] instead of [vʌdʲi̱ˈnoi̯], which is typical to some regions of Russia). For obtaining the appropriate transcription, the annotators should have similar knowledge of the pronunciation standard.

Phonemic Annotation. Phonemic transcription is based on orthoepic rules as described in pronunciation dictionaries. However, these dictionaries contain single words; in connected speech the phonemic content of a word often changes as a result of assimilatory processes (compare, e.g., Russian “под̱ березой,” under the birch tree, and “под̱ пихтой,” under the fir-tree: /paḏ-biˈrʲozaij/ and /paṯ-ˈpʲixtaj/). If such phonemic changes are well described for the language, phonemic transcription may be successfully done automatically; however, the automatic transcriber will require information about prosodic boundaries and pauses, because assimilatory processes usually do not cross large prosodic boundaries.

Dealing with assimilatory process, one may face further difficulties in transcribing connected speech. In some cases, we observe speech sounds that are absent in the phonological system of the language. For example, in Russian phonological system the feature “voiced-unvoiced” is present for most articulations: /p-b/, /t-d/, /s-z/ etc. But a few phonemes do not have their voiced or unvoiced counterparts, e.g. the phoneme /ʃʲː/. Due to regressive assimilation, some contexts may cause its voiced variant — /ʒʲː/ (e.g. “плащ̱ д̱едушки,” grandfather’s cloak: [ˈplaʒ̱ʲː ˈdʲeduʃkʲi]). As a result, even very basic pronunciation rules are not described in terms of phonemes; what we work with are actually allophones, and the resulting transcription is allophonic.

It is worth noting though that such rule-based transcription is still phonological, not phonetic. The number of possible labels for allophones is just a little greater than the number of phonemes. Purely phonemic tier may be added, if necessary, and can be easily generated automatically.

Syllabic Tier. In different language and within different linguistic traditions, syllable boundaries are defined in different ways. Among the most common approaches are the principle of sonority and the distributional principle. In the tradition of the St. Petersburg phonological school, another principle is used: Russian speech is divided into open syllables. This principle was formulated after a series of speech production experiments with “delayed feedback” (also called “artificial stuttering”) performed by Chistovich and Bondarko (1963). The researchers placed an artificial palate into the experiment participants’ mouths that prevented them from feeling their own articulation properly; at the same time, the subjects wore headphones in which their own speech was played with a significant delay. It turned out that speakers never made a break within consonant + vowel sequences, while coda consonants can be separated and form a new syllable, often with addition of a neutral vowel.

With time, the principle of open syllable acquired a number of exceptions. Among the most notable ones are syllables ending in vocalized consonant: e.g. Russian “майка,” vest: [ˈmai̯-kʌ]. In this case the division into open syllables would have produced the syllable [i̯kʌ] which, if played back on its own, is perceived as two syllables instead of one. Another exception is coda consonants at ends of large prosodic units: such sounds may be counted as quasi-syllables. If required, syllable boundaries may be automatically shifted with regard to other syllabification principles.

Word Tier. In languages like Russian, textual form of a phrase is not easily matched with real pronunciation, especially concerning stress placement. For instance, a prepositional phrase is usually pronounced with only one stress, but not necessarily; it depends on the preposition itself, on the logical structure of the phrase, and other factors. As a result, the annotation is typically performed at two different tiers: the tier for *orthographic words* (space separated) and the tier of *phonetic words* — one or more lexical words united by a single stress (e.g. Russian “не де́лали бы,” wouldn’t be doing [3rd person, plural]). Among other reasons, the tier of phonetic words is crucial for intonational research as many intonational events are anchored to the stressed syllables.

If a language allows secondary stress, a number of additional segmentation rules are required. A possible solution is to label the main stress that forms the phonetic words, as well as a few additional grades of weaker stress. The problematic cases are not only compound words, but also pronouns, conjunctions and other words that are flexible in terms of stress placement. A curious example from Russian is the pronunciation of the conjunction “но” (but), which often functions as a proclitic, but always preserves its vowel quality ([no]) — despite the fact that Russian /o/ in unstressed positions is reduced to /a/. In practice, no solution is perfect, as adding more types of stress decreases the inter-annotator agreement and reduces the accuracy of automatic transcription.

Intonation. The traditional descriptions of Russian sentence prosody are similar to those of the British School (as in O'Connor; Arnold, 1973). The basic segmental unit is an intonational phrase (IP), such that:

- an IP usually contains one main word (the nucleus) around which a linguistically relevant melodic curve is realized;
- an IP is perceived as a whole in terms of melody, tempo, loudness and pause locations;
- an IP cannot have more than one nucleus (except for specific melodic patterns where two nuclei are obligatory);
- certain prosodic phenomena occur at IP boundaries (e.g., pre-boundary lengthening).

This definition is highly suitable for prepared speech and requires further specification for disfluent speech, where IPs are often unfinished or contain an internal paralinguistic element, e.g. a filler or a silent break. As a result, we have to admit that in IP does not necessarily have to contain a nucleus (unfinished IPs), while IP-internal paralinguistic elements may induce some boundary phenomena in the middle of an IP.

Prosodic annotation must be performed by professional phoneticians based on auditory and instrumental analysis. Typically, such work requires pre-training to enhance inter-annotator agreement. In most cases, prosodic labels are added to the orthographic transcription. In all prosodically annotated *corpora*, the annotation includes the following prosodic information: boundaries of intonational phrases (IPs); location of the main word within the IP (containing the nucleus); type of melodic movement for the IP; words bearing additional prosodic prominence. Apart from the nucleus, some other syllable may bear additional prosodic prominence as perceived by the annotator. Such prominence may be manifested by any kind of prosodic parameters—notable changes in fundamental frequency (F_0), intensity, duration, voice quality, or combinations of these (Volskaya; Kachkovskaia, 2016).

Boundaries of Pitch Periods. For some purposes, we may need a precise description of the melodic contour. Automatic pitch detection algorithms sometimes make mistakes (e.g., doubling/halving errors) which result in incorrect calculations of the main prosodic parameters. For an experienced phonetician, though, it is easy to notice such errors and correct them, if the software includes that option. This can be performed by combining auditory analysis of the recording and visual analysis of the speech wave. To get the accurate F_0 values, one needs to hand-label boundaries of pitch periods; this stage is usually preceded by automatic labelling, which contains some errors.

Depending on the linguistic tradition, pitch periods can be labelled relative to different starting points (e.g., at amplitude peaks in Praat). Within the St. Petersburg phonological school, each period starts at the place of zero amplitude, where the values change from negative to positive. This is motivated by the fact that this point corresponds to the onset of the first formant (Skrelin, 1999). In general, the choice of starting point does not seem to have much influence on the resulting F_0 values.

Speech fragments produced in creaky voice are another source of pitch detection errors due to irregular vocal fold vibration and, as a result, highly variable duration of adjacent pitch periods. In such cases even manual segmentation would not enable to get the precise F_0 values. This is why at this tier, such fragments remain unannotated. If necessary, creaky voice labels may be to a special annotation tier along with other phonetic settings.

Disfluencies and Non-Speech Events. Often in non-prepared speech and rarely in prepared speech, the flow of words gets interrupted by silent breaks, filled breaks, non-phonemic sound elongations, laughs, coughs, casual non-phonemic clicks, etc. In addition, any record may contain non-speech events caused by the recording equipment or post-processing software. For successful automatic alignment of transcription with speech signal, all these events require special labelling.[4] Such labelling can be also useful for those researchers who are interested in these particular phenomena.

Setting the boundaries of these fragments is fraught with difficulties. Some of these events may occur simultaneously with speech, e.g. laughter, cough, or technical noises. This means that their boundaries must be labelled on an additional annotation tier, or even several additional tiers. Another problem concerns fillers (such as “ehm” and

[4] From our experience, reflecting speech disfluencies in the orthographic transcription may add up to 10% to the total number of speech sounds (Kachkovaskaia et al., 2016).

"uhm"): when such element begins right after a vowel phoneme, especially an open vowel, it is hard to detect the boundary between the vowel phoneme and the filler. For solving most tasks, however, precise labelling of these events is unnecessary. This is why the optimal decision is to annotate them within one of the other tiers, e.g. at the intonational tier. Some of these events may also be labelled automatically, most particularly silent breaks.

3 Annotated Speech *Corpora* Developed at the Department of Phonetics

3.1 The INTAS *Corpus* of Russian Speech

The annotation and segmentation principles formulated above can be best illustrated by the *INTAS* Corpus *of Russian Speech*. It was created during INTAS project 915 *Spontaneous speech of typologically unrelated languages: Russian, Finnish and Dutch* (Skrelin, 2009). For this *corpus*, ten native Russian speakers (5 male, 5 female) were recorded. The speakers represented different age groups (< 20, 20–30, 30–40, 40–50, and > 50), and each speaker used the St. Petersburg variant of the Russian pronunciation standard.

The first stage consisted in recording informal dialogues between familiar speakers. From each recording, a 5-minute fragment was selected for further analysis. At the second stage, each speaker was asked to read a text constructed from his own monologue.

Segmentation was performed by different phoneticians who followed the segmentation rules used for concatenation of allophones in the allophone-based speech synthesis system (Skrelin, 1999). During the segmentation, preliminary segmental transcription was performed. At the same time, pitch was automatically detected and manually corrected. Segmentation and transcription were then checked and corrected by two experienced phoneticians.

The annotation scheme contains 8 tiers (see Fig. 1):

1. acoustic: phonetic transcription (segmentation and labeling) produced by listening to the isolated speech sounds using instrumental methods (spectrograms);

2. perceptual: phonetic transcription (segmentation and labeling) produced by listening to sounds within the word;

3. phonemic (ideal): phonemic transcription (segmentation and labeling) based on the standard Russian pronunciation rules;

4. syllabic: segmentation into open syllables using the transcription on tier 1;

5. stress: indication of lexical stress;

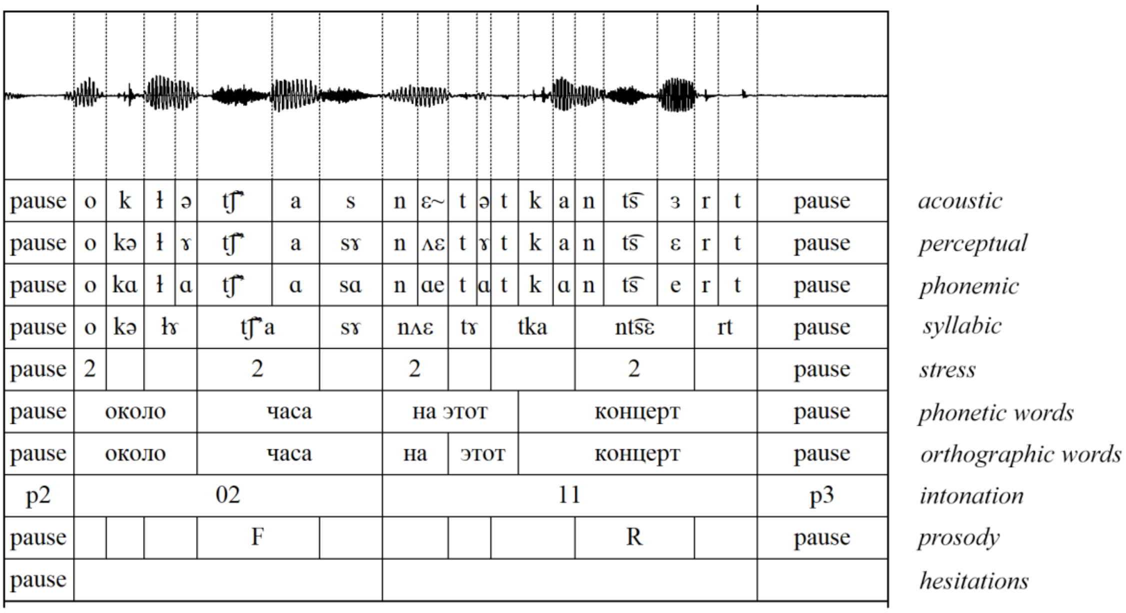
6. phonetic words: content words and their surrounding clitics;

7. orthographic words: space-separated words;

8. intonation units: segmentation and labeling of intonational phrases, with information on the melodic type for each IP and type of pause according to the annotation system suggested by N. Volskaya (Volskaya, Kachkovskaia, 2016);

9. prosody: labelling of the main melodic movements (rising vs. falling);

10. hesitation events (filled pauses): segmentation and labeling.



pause	o	k	ł	ə	t͡ʃ	a	s	n	ɛ~	t	ə	t	k	a	n	t͡s	ʒ	r	t	pause	*acoustic*
pause	o	kə	ł	ɤ	t͡ʃ	a	sɤ	n	ʌɛ	t	ɤ	t	k	a	n	t͡s	ɛ	r	t	pause	*perceptual*
pause	o	kɑ	ł	ɑ	t͡ʃ	ɑ	sɑ	n	ɑe	t	ɑ	t	k	ɑ	n	t͡s	e	r	t	pause	*phonemic*
pause	o	kə	łɤ		t͡ʃa		sɤ	nʌɛ		tɤ		tka			nt͡sɛ			rt		pause	*syllabic*
pause	2				2			2							2					pause	*stress*
pause	около				часа			на этот					концерт							pause	*phonetic words*
pause	около				часа			на		этот			концерт							pause	*orthographic words*
p2	02							11												p3	*intonation*
pause					F										R					pause	*prosody*
pause																					*hesitations*

Fig. 1. Annotation tiers in the INTAS speech *corpus*

Figure 1 shows an example of annotation performed in Praat for the phrase “[времени было потрачено] около часа на этот концерт” ([we spent] around an hour on this concert). Comparing the acoustical and perceptual tiers, we may find cases of vowel omissions in real pronunciation. The syllabic tier clearly demonstrates the principle of the open syllable; there is also an example of resyllabification (“этот концерт,” syllable [tka]) which defies the strict layering principle. At the intonation tier, we may notice two IPs[5] with no pause between them, and pauses to the left and to the right; the prosodic tier helps us see where exactly the melodic movements (the fall, labelled “F,” and the rise, labelled “R”) are located. The hesitation tier is empty in this example, but in other parts of the recordings it marks the boundaries of filled pauses.

3.2 Other *Corpora*

The information about other *corpora* mentioned in this paper is summarized in Tables 1 and 2. Below, we will provide brief descriptions of the most notable speech *corpora* developed at the Department of phonetics. This review does not include specialized databases of sound segments that were developed for automatic speech synthesis systems for Russian (Skrelin, 1997a).

Corpus of Russian Professionally REad Speech (CORPRES) (Skrelin et al., 2010) was created in 2009–2011 and originally intended for use in unit-selection text-to-speech synthesis. However, being a good representation of standard Russian speech, it has been used in a large number of phonetic research projects. Recordings were made from professional[6] speakers with the St. Petersburg variant of the Russian pronunciation standard.

Corpus of Russian Spontaneous Speech (CoRuSS) (Kachkovskaia et al., 2016). The main purpose of the work was to create a database of non-read speech passages recorded from speakers of different age and gender groups. The *corpus* was intended to be used for research in automatic prosodic boundary detection. The recordings were made in the form of spontaneous dialogues. Apart from spontaneous dialogue, each speaker also recorded the phonetically representative text and a short monologue about himself/herself.

During the creation of this *corpus*, the annotation system was further elaborated to include various kinds of disfluencies and paralinguistic events. Below one may find a fragment from the orthographic tier containing prosodic annotation along with many

[5] Melodic type 02 is a (rising-)falling tonal movement within the nucleus often used for emphasis or contrast; type 11 is a rising(-falling) tonal movement often used in non-final IPs. Pause type p2 corresponds to a weaker prosodic break than p3.

[6] Mostly, TV and radio broadcasters.

other phenomena. In this example,[7] slashes correspond to IP boundaries, [02], [09], [11] and [11b] represent the melodic types (tunes), [+] marks additional prosodic prominence, “9” corresponds to laughter, “э-” is a vocalic hesitation of any quality, labels “1” and “2” after vowels are used to mark strong and weak stress, respectively, and colons mark elongations of speech sounds.

> [02]не1 / ну [+]ла1дно та1м [02]преподава1тели / я2 ду1маю что2 они2 таки1е у на2с лю1ди [09]обеспе1ченные / 9 / а [11b]аспира1нтам / 9 / э- ну1 в осо1бенности ка1к [02]лё1ха / 9 / и1м оставля1ют то1лько [11]с:типе1ндию / кото1рая [02]госуда1рственная /

SibLing Corpus of Russian Dialogue Speech (Kachkovskaia et al., 2020). This *corpus* was developed specifically for research on speech entrainment—the phenomenon of speakers’ attuning to each other during conversation, which results in similarities in interlocutors’ gestures, mimics, and speech. In SibLing, the basic set of speakers were 10 pairs of same-gender siblings aged 23 to 40. Each of these 20 speakers communicated 5 different interlocutors (invited speakers) with varying degree of familiarity and “social distance”: from siblings or close friends to strangers of significantly greater age. During the recording each pair of interlocutors performed two collaborative tasks: a card-matching game and map task.

Multimedia Corpus of Ironic Speech (Kochetkova et al., 2021) was developed in the framework of the project *Acoustic correlates of irony with respect to basic types of pitch movement.* The *corpus* contains reading of 330 short monologues and dialogues, as well as four long coherent texts, which included homonymous ironic and non-ironic utterances of various communicative types enabling implementation of all possible melodic patterns. The required connotations (ironic vs. non-ironic) were stimulated by the context: by means of lexical, grammatical or semantic markers of irony, as well as by context only. Along with the audio, the *corpus* contains high-speed video recordings.

Title	**Material**	**Speakers**	***Corpus* size**	**Availability**
INTAS *Corpus* of Russian Speech	5 minutes of spontaneous speech + 5 minutes of reading (a text with	10 speakers, different	1h 40 min	Available[8]

[7] In English: [02]no1 / well [+]oka1y tho1se [02]te1achers / i2 thi1nk tha2t the2y a2re ki1nd of pe1ople [09]be2tter-o1ff / 9 / and [11b]stu1dents / 9 / э- we1ll espe1cially li1ke [02]a1lex / 9 / the1y a2re le1ft with o1nly [11]s:cho1larship / whi1ch is [02]bu1dgetary /

[8] henceforth: available for academic purposes on demand (by contacting the creators)

	approx. the same lexical content)	gender and age groups		
CORPRES	Fictional and non-fictional texts read by professional speakers	4 males, 4 females	60 h	Unavailable[9]
CoRuSS	Spontaneous dialogues (free conversation)	60 speakers, different gender and age groups	30 h	Available
SibLing	Cooperative dialogues (map task and card-matching game)	100 speakers	64 h	Available
Multimedia *Corpus* of Ironic Speech	Reading short monologues and dialogues, and long coherent texts (incl. homonymous ironic and non-ironic utterances of various communicative types)	56 speakers, different gender and age groups	12 h	Available

Table 1. Speech *corpora* developed recently at the Department of Phonetics, St. Petersburg State University.

Tier	*Corpora*				
	INTAS	CORPRES	CoRuSS	SibLing	Ironic speech
Acoustic phonetic	S(m), L(m)	S(m), L(m)			
Perceptual phonetic	S(m), L(m)				
Phonemic	S(m), L(m)	S(m), L(m)	L(a)	L(a)	
Syllabic	S(m), L(m)				
Word	S(m), L(m)	S(m), L(m)	L(m)	L(m)	L(m)
Word stress	S(m), L(m)	S(m), L(m)	L(m)	L(m)	
Intonation	S(m), L(m)	S(m), L(m)	L(m)	L(m)	L(m)
Pauses	S(m), L(m)	S(m), L(m)	S(m), L(m)	S(m), L(m)	
Pitch periods	S(m), L(m)	S(m), L(m)		S(a), L(a)	
Disfluencies	S(m), L(m)		S(m), L(m)	L(m)	
Non-speech events			S(m), L(m)	L(m)	

Table 2. The annotation scheme of speech *corpora* developed recently at the Department of Phonetics, St. Petersburg State University: S – segmented, L – labelled; segmentation and labelling were either (m)annual or (a)utomatic.

[9] The corpus belongs to a commercial company.

Speech Databases and Collections. There are also a number of speech databases and speech collections which were annotated partially following the principles describes in Section 2:

- the database of speech recordings by V. M. Zhirmunsky (Svetozarova, 1996);
- Tales of the Russian North (Skrelin et al., 1997b);
- Poetic Folklore of the Russian North (Lamentations) (Skrelin, 1998);
- Russian speech of Canadian doukhobors (Makarova et al., 2011);
- speech recordings for vocal fatigue research (Evgrafova et al., 2016);
- recordings of professional singers (Evdokimova et al., 2017).

4 Speech *Corpora* in Phonetic Research

Many years of experience in collecting, processing and analysing speech material have enabled us to create speech *corpora* of all kinds that can serve as the basis for a wide range of fundamental and applied research. The fully annotated large *corpus* of read speech CORPRES laid the foundation for a lot of research projects including automatic prosodic boundary detection (Kocharov et al., 2019a), research on vowel reduction (Kocharov et al., 2019b) and phrase-final lengthening (Kachkovskaia et al, 2013), melodic declination (Kocharov et al., 2015), the melody of post-nucleus (Kachkovskaia et al., 2020) and others.

The two large annotated speech *corpora* CORPRES and CoRuSS were used as material for comparing read speech and spontaneous speech in terms of intonational phrase duration and length, distribution of melodic types, silent pause duration, frequency of marking IP boundaries by real silent pauses (Kachkovskaia, Skrelin, 2020).

The SibLing *corpus* serves as the major source of speech data for research on speech entrainment (Menshikova et al., 2020). The specific design of this *corpus* enables to trace the influence of social and situational factors on the interlocutors' speech (Kachkovskaia et al, 2022).

The Multimedia *corpus* of ironic speech served as a base for the comparison of ironic speech phonetic features in Russian vs. French languages (Skrelin et al., 2021), perception of irony in male vs. female speech (Kochetkova et al., 2020).

The research on acoustic cues of vocal fatigue was performed on recordings of speakers before and after the vocal load (Evdokimova et al., 2017). The research on acoustic cues of voice pathologies in singing speech was performed using recordings of professional singers (Evdokimova et al., 2019).

Conclusion

In phonetic research, it is often crucial to analyse complex interaction between factors functioning at different levels. This is why one of the major principles that we follow is to analyse the speech material at all levels. Thus, a fully annotated speech *corpus* would include the following annotation tiers that reflect the main principles of *corpus* design based on ideas of the St. Petersburg phonological school.

1. Phonetic tier 1 (acoustic). The major principles: (1) de-lexicalization and (2) taking into account the acoustic properties of sounds.
2. Phonetic tier 2 (perceptual) produced by listening to sounds within the word.
3. Phonemic tier based on standard pronunciation rules.
4. Syllabic tier which reflects resyllabification.
5. Phonetic words (clitic groups) tier, with stress markings.
6. Intonation tier produced by experts based on auditory and instrumental analysis.
7. Pitch tier: produced automatically with subsequent manual correction.
8. Speech disfluencies and non-speech events: produced manually to reflect technical noises, hesitations, false-starts, elongations, laughter etc.

This annotation scheme is very time-consuming, especially if a large dataset is required. Given a specific research task, we may omit some of these tiers. Intonation research, e.g., would not necessarily require full segmental annotation: segmental units are only needed to calculate peak alignment, but for this we would only need boundaries of accented vowels, but not all speech segments. The *corpus* of spontaneous speech CoRuSS was developed for research in automatic prosodic boundary detection, and thus does not include boundaries of speech sounds. However, it still contains orthographic and phonetic transcription—which means that we might be able to add the missing tiers later, when the algorithms for automatic speech alignment are able to provide higher accuracy than nowadays. Our latest attempts provided rather high error (on average, around 20 ms), but as soon as we reach significantly lower numbers (at least around 6 ms), we will get the opportunity to add high-quality segmental tiers to those *corpora* where these tiers are missing.

REFERENCES

Prilozhenie №3 k Byulletenyu Foneticheskogo fonda russkogo yazyka. Fond zvukovyh edinits russkoi rechi [Appendix # 3 to the Bulletin of the Russian Phonetic Fund]. Russian Phonetic Fund. St. Petersburg - Bochum, 1993.

BONDARKO, L. V.; SVETOZAROVA, N. D.; SKRELIN, P. A. *Foneticheskii fond russkogo yazyka kak issledovatel'skaya programma kafedry fonetiki Leningradskogo universiteta* [Russian Phonetic Fund as a Research Program of Department of Phonetics, Leningrad University]. Byulleten' Foneticheskogo fonda russkogo yazyka, St. Petersburg - Bochum, n.4, 1992.

BONDARKO, L. V.; VERBITSKAYA, L. A. (ed.) *Interferenciya zvukovyx sistem* [Cross-Language Influence of Sound Systems], Leningrad: Izdatel'stvo LGU, 1987.

BONDARKO, L. V.; VERBITSKAYA, L. A.; GORDINA, M. V.; KASEVICH, V. B. Stili proiznosheniya i tipy proizneseniya [Styles and Types of Pronunciation]. *Voprosy yazykoznaniya*, Moscow, n. 2. pp.64-70, 1974.

CHISTOVICH, L. A.; BONDARKO, L. V. Ob upravlenii artikulyatsionnymi organami v processe rechi [About Controlling Articulatory Organs in Speech Production]. *In: Issledovalia po strukturnoj tipologii* [Research in Structural Typology]. Moscow: Nauka, pp.169-182, 1963.

CHODROFF, E.; WILSON, C. Structure in Talker-Specific Phonetic Realization: Covariation of Stop Consonant VOT in American English. *Journal of Phonetics*, v. 61, pp.30-47, 2017.

EVDOKIMOVA, V.; EVGRAFOVA, K.; CHUKAEVA T. The Database of Normal and Pathological Singers' Voices: An Approach to Collecting Data. *In*: The 10th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), 10., 2017, Florence. *Proceedings* [...]. Florence: Firenze University Press, 2017. pp.23-24.

EVDOKIMOVA, V.; KOCHAROV, D.; SKRELIN, P. Method for Constructing Formants for Studying Phonetic Characteristics of Vowels. *SPIIRAS Proceedings*, v. 19(2), pp.302-329, 2020.

EVDOKIMOVA, V.; SKRELIN, P.; CHUKAEVA, T. Automatic Phonetic Transcription for Russian: Speech Variability Modeling. *In:* International Conference on Speech and Computer (SPECOM), 19, 2017, Hatfield. *Proceedings* [...]. Springer International Publishing, 2017. pp.192-199.

EVDOKIMOVA, V.; ZAKHARCHENKO, E.; SKRELIN, P.; EVGRAFOVA, K.; CHUKAEVA, T.; SHVALEV, N. Akusticheskie xarakteristiki golosa v rechi i penii opernyx pevczov v norme i pri patologii [Acoustic Characteristics of Voice in Speech and Singing of Opera Singer's for Normal and Pathological Voice]. *In*: Interdisciplinary

Seminar on Conversational Russian Speech Analysis, 8., 2019, Saint Petersburg. *Proceedings* […], Saint Petersburg: Polytechnika-print, 2019. pp.21-30.

EVGRAFOVA, K.; EVDOKIMOVA, V.; CHUKAEVA, T.; SKRELIN, P. Vocal Fatigue in Voice Professionals: Collecting Data and Acoustic Analysis. *In*: Tutorial and Research Workshop on Experimental Linguistics (EXLING 2016), 7., 2016, Saint-Petersburg. *Proceedings* [...], Saint-Petersburg: Saint Petersburg State University, 2016. pp.59-62.

GAROFOLO, J.; LAMEL, L.; FISHER, W.; FISCUS, J.; PALLETT, D.; DAHLGREN, N.; ZUE, V. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993.

KACHKOVSKAIA, T.; CHUKAEVA, T.; EVDOKIMOVA, V.; KHOLIAVIN, P.; KRIAKINA, N.; KOCHAROV, D.; MAMUSHINA, A.; MENSHIKOVA, A.; ZIMINA, S. SibLing Corpus of Russian Dialogue Speech Designed for Research on Speech Entrainment. *In*: Conference on International Language Resources and Evaluation (LREC 2020), 12., Marseille. *Proceedings* [...], Marseille: ELRA, 2020. pp.6556-6561.

KACHKOVSKAIA, T.; KOCHAROV, D.; SKRELIN, P.; VOLSKAYA, N. CoRuSS—A New Prosodically Annotated Corpus of Russian Spontaneous Speech. *In*: Conference on International Language Resources and Evaluation (LREC 2016), 10., 2016, Portorož. *Proceedings* [...], Portorož: ELRA, 2016. pp.1949-1954.

KACHKOVSKAIA, T.; MAMUSHINA, A.; PORTNOVA, A. Typical and Rare Post-Nuclear Melodic Movements in Russian. *In*: Speech Prosody, 10., 2020, Tokyo. *Proceedings* [...], Tokyo: ISCA, 2020. pp.464-468.

KACHKOVSKAIA, T., MENSHIKOVA, A.; KOCHAROV, D.; KHOLIAVIN, P.; MAMUSHINA, A. Social and Situational Factors of Speaker Variability in Collaborative Dialogues. *In*: Speech Prosody, 11., 2022, Lisbon. *Proceedings* [...], Lisbon: ISCA, 2022. pp.455-459

KACHKOVSKAIA, T.; SKRELIN, P. Prosodic Phrasing in Russian Spontaneous and Read Speech: Evidence from Large Speech Corpora. *In*: Speech Prosody, 10., 2020, Tokyo. *Proceedings* [...], Tokyo: ISCA, 2020. pp.166-170.

KACHKOVSKAIA, T.; VOLSKAYA, N.; SKRELIN, P. Final Lengthening in Russian: A Corpus-Based Study. *In*: Interspeech 2013, 14., Lyon. *Proceedings* [...], Lyon: ISCA, 2013. pp.1438-1442.

KOCHAROV, D.; KACHKOVSKAIA, T.; SKRELIN, P. Prosodic Boundary Detection Using Syntactic and Acoustic Information. *Computer Speech and Language*, v. 53, pp.231-241, 2019a.

KOCHAROV, D.; KACHKOVSKAIA, T.;SKRELIN, P. Prosodic Factors Influencing Vowel Reduction in Russian. *In*: Interspeech 2019, 20., Graz. *Proceedings* [...], Graz: ISCA, 2019b. pp.1956-1960.

KOCHAROV, D.; VOLSKAYA, N.; SKRELIN, P. F0 Declination in Russian Revisited. *In*: International Congress of Phonetic Sciences (ICPHS), 18., 2015, Glasgow. *Proceedings* [...], Glasgow: International Phonetic Association, 2015.

KOCHETKOVA, U.; SKRELIN, P.; EVDOKIMOVA, V.; NOVOSELOVA, D. Perception of Irony in Speech. *In*: International Conference on Neurobiology of Speech and Language, 4., 2020, Saint Petersburg. *Proceedings* [...], Saint Petersburg: Skifia-Print, 2020. pp.72-73.

KOCHETKOVA, U.; SKRELIN, P.; EVDOKIMOVA, V.; NOVOSELOVA, D. The Speech Corpus for Studying Phonetic Properties of Irony. *In*: Language, Music and Gesture: Informational Crossroads, 2021, Saint Petersburg. *Proceedings* [...], Springer International Publishing, 2021. pp.203-214.

LIBERMAN, M. Corpus Phonetics. *Annual Review of Linguistics*, 5, pp.91-107, 2019.

MAKAROVA, V. A.; USENKOVA, E. V.; EVDOKIMOVA, V. V.; EVGRAFOVA, K. V. Yazyk saskachevanskix duxoborov: vvedenie v analiz [The Language of the Saskatchevan Doukhobors: Introduction and Analysis]. *Izvestiya vysshix uchebnyx zavedenij*. Seriya «Gumanitarnye nauki». Razdel lingvistika [New of Higher School. Humanities. Linguistics], Ivanovo, v. 2, n. 2, pp.146-152, 2011.

MENSHIKOVA, A.; KOCHAROV, D.; KACHKOVSKAIA, T. Phonetic Entrainment in Cooperative Dialogues: A Case of Russian. *In*: Interspeech 2020, 21., Shanghai. *Proceedings* [...], Shanghai: ISCA, 2020. pp.4148-4152.

O'CONNOR, J. D.; ARNOLD, G. F., *Intonation of Colloquial English*. Bristol, U.K.: Longman Group Ltd., 1973.

OSTENDORF, M.; PRICE P. J.; SHATTUCK-HUFNAGEL, S., *The Boston University Radio News Corpus*, Boston University Technical Report No. ECS-95-001, 1995.

PANAYOTOV, V.; CHEN, G.; POVEY, D.; KHUDANPUR, S. Librispeech: an ASR Corpus Based on Public Domain Audio Books. *In*: 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 40., 2015, Brisbane. *Proceedings* [...], Brisbane: IEEE, pp.5206-5210.

PAPERNO, S.; LEED, R.L. Vocabulary Words in Elementary Russian Textbooks. *Slavic and Eats European languages journal*, v. 32, n. 2, 1988. pp.305-312

SKRELIN, P. Concatenative Russian Speech Synthesis: Sound Database Formation Principles. *In*: International Conference on Speech and Computer (SPECOM), 2., 1997. Cluj-Napoca. *Proceedings* [...], Cluj-Napoca: Editura Promedia Plus, 1997a.

SKRELIN, P. A. (ed.) Skazki Russkogo Severa [Tales of the North of Russia]. *In*: *Byulleten` foneticheskogo fonda russkogo yazyka. Prilozhenie 6* [The Bulletin of the Russian Phonetic Fund. Appendix 6]. Saint Petersburg - Bochum, 1997b.

SKRELIN, P. A. (ed.) Obryadovaya poeziya Russkogo Severa: plachi [Poetic Folklore of the North of Russia (Lamentations)]. *In*: *Byulleten` foneticheskogo fonda russkogo yazyka. Prilozhenie 6* [The Bulletin of the Russian Phonetic Fund. Appendix 6], Saint Petersburg - Bochum, 1998.

SKRELIN, P. A. *Segmentaciya i transkripciya* [Segmentation and Transcription], Saint Petersburg: Saint Petersburg State University, 1999.

SKRELIN, P. Russian Material and Methods. *In*: DE SILVA, V.; ULLAKONOJA, R. (ed.) *Phonetics of Russian and Finnish*, Frankfurt am Main: Peter Lang, 2009.

SKRELIN, P. A.; KOCHETKOVA, U. E.; EVDOKIMOVA, V. V.; NOVOSELOVA, D. D.; GERMAN, R. D. Prosodicheskie xarakteristiki ironicheskix vyskazyvanij v russkom i franczuzskom yazykax [Prosodic Features of Ironic Utterances in Russian and French]. *In*: Interdisciplinary Seminar on Conversational Russian Speech Analysis, 9., 2021, Saint Petersburg. *Proceedings* […], Saint Petersburg: Skifia-Print, 2021. pp.81-86.

SKRELIN, P.; VOLSKAYA, N.; KOCHAROV, D.; EVGRAFOVA, K.; GLOTOVA, O.; EVDOKIMOVA, V. A Fully Annotated Corpus of Russian Speech. *In*: Conference on International Language Resources and Evaluation (LREC 2010), 7., 2010, Valletta. *Proceedings* [...], Valletta: ELRA, 2010. pp.109-112.

SVETOZAROVA, N. Zhirmunsky's Collection of German Folk Songs in the Sound Archives of the Pushkinsky Dom. *In*: *Archives of the Languages of Russia*. Saint Petersburg - Groningen, 1996. pp.33-38.

TURK, A.; NAKAI, S.; SUGAHARA, M. Acoustic Segment Durations in Prosodic Research: A Practical Guide. Methods. *In*: *Empirical Prosody Research*, Berlin, Boston: De Gruyter, pp.1-28, 2012.

VOLSKAYA, N.; KACHKOVSKAIA, T. Prosodic Annotation in the New Corpus of Russian Spontaneous Speech CoRuSS. *In*: Speech Prosody, 8., 2016, Boston. *Proceedings* [...], Boston: ISCA, 2016. pp.917-921.

WANG, C.; RIVIERE, M.; LEE, A.; WU, A.; TALNIKAR, C.; HAZIZA, D.; WILLIAMSON, M.; PINO, J.; DUPOUX, E. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. *In*: ACL 2021 (Volume 1: Long Papers), Bangkok, *Proceedings* […], Bangkok: ACL, 2021. pp.993-1003.

Received September 30, 2021

Accepted March 20, 2023

Declaration of Author's Contribution

The author's contributions are stated to be the following:
Pavel Skrelin: Conceptualization of initial principles to design speech corpora; Evolution of the methodology to design speech corpora; Writing original draft and article revision; Final revision and approval for the publishing.
Tatiana Kachkovskaia: Evolution of the methodology to design speech corpora; Writing original draft, article revision and editing; Final revision and approval for the publishing.
Daniil Kocharov: Evolution of the methodology to design speech corpora; Writing original draft, article revision and editing; Final revision and approval for the publishing.
Vera Evdokimova: Evolution of the methodology to design speech corpora; Final revision and approval for the publishing.
Uliana Kochetkova: Evolution of the methodology to design speech corpora; Final revision and approval for the publishing.

Research Data and Other Materials Availability

The contents underlying the research text are included in the manuscript.

Reviews

Due to the commitment assumed by *Bakhtiniana. Revista de Estudos do Discurso* [*Bakhtiniana.* Journal of Discourse Studies] to Open Science, this journal only publishes reviews that have been authorized by all involved.