CoRuSS — a New Prosodically Annotated Corpus of Russian Spontaneous Speech

Tatiana Kachkovskaia, Daniil Kocharov, Pavel Skrelin, Nina Volskaya
Department of Phonetics, Saint Petersburg State University,
St.Petersburg State University, 7/9 Universitetskaya nab., St. Petersburg, 199034 Russia
[kachkovskaia, kocharov, skrelin, volni]@phonetics.pu.ru

Abstract

This paper describes speech data recording, processing and annotation of a new speech corpus CoRuSS (Corpus of Russian Spontaneous Speech), which is based on connected communicative speech recorded from 60 native Russian male and female speakers of different age groups (from 16 to 77). Some Russian speech corpora available at the moment contain plain orthographic texts and provide some kind of limited annotation, but there are no corpora providing detailed prosodic annotation of spontaneous conversational speech. This corpus contains 30 hours of high quality recorded spontaneous Russian speech, half of it has been transcribed and prosodically labeled. The recordings consist of dialogues between two speakers, monologues (speakers' self-presentations) and reading of a short phonetically balanced text. Since the corpus is labeled for a wide range of linguistic—phonetic and prosodic—information, it provides basis for empirical studies of various spontaneous speech phenomena as well as for comparison with those we observe in prepared read speech. Since the corpus is designed as a open-access resource of speech data, it will also make possible to advance corpus-based analysis of spontaneous speech data across languages and speech technology development as well.

Keywords: speech corpus, speech annotation, Russian

1. Introduction

At the moment a number of medium and large size Russian speech corpora are available.

The largest published corpus of Russian speech is ORD (One Day of Speech) corpus that is still under development (Bogdanova-Beglarian et al., 2015). It contains more than 1000 hours of everyday speech. The corpus is collected by recording the whole day speech of more than 100 participants. It has partial annotation and transcription. However, this corpus is not publicly available.

The most annotated publicly available corpus nowadays is PrACS-Russ (Prosodically Annotated Corpus of Spoken Russian), which contains over 4 hours of monologue speech (Kibrik and Podlesskaya, 2009) with limited prosodic annotation based on two types of tones—rise and fall—and their combinations. The annotation also includes a detailed labelling of various speech phenomena such as speech errors, self-corrections, filled pauses. It is available as part of Russian National Corpus (Apresjan et al., 2006). The corpus may be a valuable source for research of discourse events in monologue speech, but rather poor recording quality makes it possible to use it only for basic phonetic research. Due to inconsistent and variable noise, these recordings are not suitable for acoustic analysis and modelling in the field of speech technologies.

The corpora containing well-annotated high-quality recordings are not publicly available. It is worth mentioning some of them for their well-structured design and annotation.

Corpus of Professionally Read Speech (CORPRES) contains over 30 hours of speech recorded in a professional studio (Skrelin et al., 2010). Segmented and annotated on orthographic, phonetic, and prosodic levels, it contains a manually corrected pitch tier and information about various phonetic phenomena such as epenthetic vowels, laryngealization, and glottalization. It is a rich source of data for any kind of phonetic research on read speech and individual speech production strategies. However, there are only eight speakers recorded.

Corpus of monologues "RuSpeech" contains about 50 hours of transcribed recordings produced by 220 speakers (Krivnova, 2013). The speaker variability and high-quality recordings allow to use it for training acoustic models for ASR or TTS.

CoRuSS (Corpus of Russian Spontaneous Speech) presented in this paper is designed as a publicly available resource containing high-quality recordings of spontaneous speech with detailed prosodic transcription. The recordings include dialogues between native Russian speakers, with a part of it—at least 14 hours of speech from 60 speakers—annotated by expert linguists at lexical and prosodic levels.

2. Corpus design and creation

There are three types of recordings within the corpus:

- free conversation between two speakers (dialogue),
- speaker's introduction (monologue),
- reading the phonetically balanced text.

2.1. Speakers

The recordings were made from 60 native Russian speakers of three age groups, with 10 males and 10 females in each group:

- 16 to 30
- 31 to 45
- 46 to 77.

The speakers were asked to fill in the form providing information about their age, sex, education, profession, birthplace, native language, experience in pronunciation practice of foreign languages, cities where he/she had attended school and college/university, cities where he/she had lived

Table 1: Speaker information: gender, age and duration of transcribed and annotated dialogue (minutes:seconds)

ID	Gender	Age	Dialogue	ID	Gender	Age	Dialogue	ID	Gender	Age	Dialogue
M01	M	38	13:31	M21	M	32	11:45	M41	M	59	12:19
M02	M	37	11:36	F22	F	24	12:11	M42	M	26	11:04
M03	M	37	10:15	F23	F	34	13:08	M43	M	30	22:12
M04	M	34	15:02	F24	F	36	14:54	F44	F	32	14:13
M05	M	60	13:25	F25	F	45	13:02	M45	M	51	12:38
M06	M	23	19:49	F26	F	37	11:53	F46	F	35	12:07
M07	M	23	16:23	M27	M	32	12:36	F47	F	21	11:57
F08	F	24	10:01	M28	M	31	15:21	F48	F	28	12:34
M09	M	51	14:05	M29	M	20	12:27	F49	F	31	18:59
F10	F	17	12:01	F30	F	60	13:48	M50	M	50	12:16
F11	F	19	12:25	F31	F	63	12:30	M51	M	54	12:46
M12	M	24	22:26	F32	F	69	19:23	M52	M	53	12:26
M13	M	24	28:20	F33	F	77	15:17	M53	M	65	11:38
F14	F	18	13:53	F34	F	60	14:39	F54	F	54	14:33
M15	M	19	19:01	F35	F	58	20:16	F55	F	53	13:59
F16	F	23	16:03	F36	F	40	14:11	M56	M	32	16:45
M17	M	25	16:37	M37	M	16	12:19	F57	F	41	11:26
M18	M	26	18:18	F38	F	68	11:55	M58	M	68	12:11
M19	M	41	12:10	M39	M	69	12:05	F59	F	44	12:35
F20	F	19	23:37	F40	F	47	13:57	F60	F	24	12:55

for at least one year. They were also asked to define their general physical and emotional state at the time of the recording.

All the speakers mentioned Russian as their only native language. Of the 60 speakers, 51 had higher education, 7 were students, and 2 had only finished high school (their age was below 20). 42 had attended school in Saint Petersburg, and 18—in other regions of Russia and the former Soviet Union. 49 mentioned their previous experience in pronunciation practice of foreign languages. 31 speaker defined their state at the time of the recording as ‘normal’, 23 as ‘very good’, 5 as ‘tired’, and 1 as ‘feeling unwell’.

In most cases the two speakers were friends or relatives and knew each other very well. However, some participants only met for the first time in the studio.

When the speaker came to the studio alone and no other participant could be found at the time, a staff member became his dialogue partner, whose speech was not recorded for the corpus. In total 14 speakers were recorded in this manner.

2.2. Recording set-up

The speakers sat in the soundproof studio opposite each other at a comfortable distance of 1.5–2 metres. The recording equipment consisted of MOTU Traveler FireWire audio interface and microphones. Each speaker was recorded using AKG HSC 271, an individual headset equipped with a condenser microphone with cardioid polar pattern. Additionally, a bi-directional microphone was placed between the speakers (Audio-Technica AT 2050, a condenser microphone with figure-8 polar pattern). The sampling rate was 44100 Hz, bit rate—16 bits.

Thus, speech was recorded from three sources in multichannel mode. The recorded speech was exported into three separate audio files.

Since the speakers were in the same room, the signal from the headsets inevitably contains speech of both of them; the ratio between the speaker’s and his/her interlocutor’s intensity is around 11 dB.

Two staff members managed the recording procedure: a sound engineer and a supervisor. The task of the supervisor was to interfere in the conversation if (and only if) one of the speakers spoke too little, and encourage him to get more involved.

2.3. Recording procedure

The main part of the recording session was a conversation between two speakers, which lasted from 30 to 70 minutes depending on the quality of recording and speech material. The goal was to obtain at least 10 minutes of clear speech per speaker. The conversation was natural, the range of topics was not limited in any way. The speakers talked about traveling, family, science, education, personal life etc.

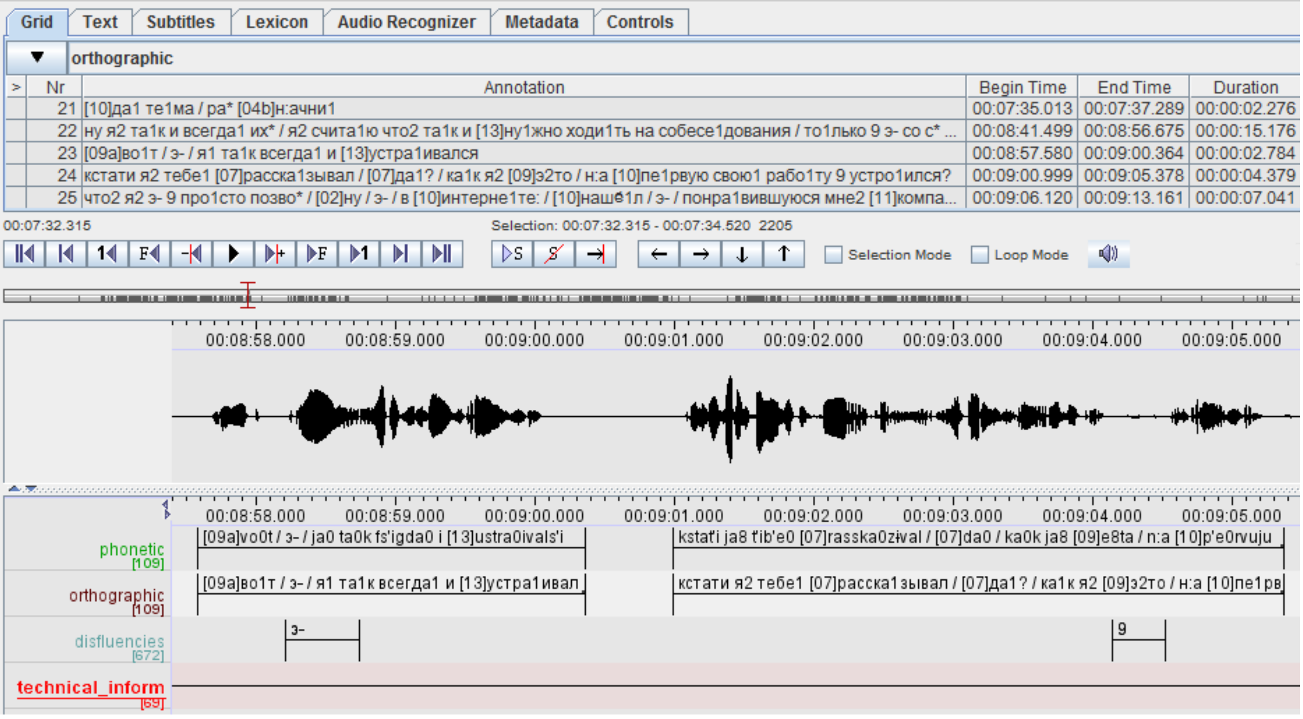
When the dialogue was over, the speakers were asked to introduce themselves and afterwards read a phonetically balanced text; for this part the speaker was in the studio alone. Due to personal reasons, 7 of the 60 speakers could not participate in this part of the recording.

Personal monologues took 0.5 to 3 minutes depending on the speaker. The total duration of all monologues is 52 minutes.

Reading of the phonetically balanced text, 438 words long, took 2.5 to 4.5 minutes. The text was the same for all speakers. The total duration of read speech is 3 hours.

Prior to annotation, the recordings were processed automatically in order to identify defective speech fragments so as

Figure 1: An example of corpus annotation in ELAN



Nr	Annotation	Begin Time	End Time	Duration
21	[10]да1 те1ма / ра* [04b]н:ачни1	00:07:35.013	00:07:37.289	00:00:02.276
22	ну я2 та1к и всегда1 их* / я2 счита1ю что2 та1к и [13]ну1жно ходи1ть на собесе1дования / то1лько 9 э- со с* ...	00:08:41.499	00:08:56.675	00:00:15.176
23	[09a]во1т / э- / я1 та1к всегда1 и [13]устра1ивался	00:08:57.580	00:09:00.364	00:00:02.784
24	кстати я2 тебе1 [07]расска1зывал / [07]да1? / ка1к я2 [09]э2то / н:а [10]пе1рвую свою1 рабо1ту 9 устро1ился?	00:09:00.999	00:09:05.378	00:00:04.379
25	что2 я2 э- 9 про1сто позво* / [02]ну / э- / в [10]интерне1те: / [10]наше1л / э- / понра1вившуюся мне2 [11]компа...	00:09:06.120	00:09:13.161	00:00:07.041

phonetic [109]: [09a]vo0t / э- / ja0 ta0k fs'igda0 i [13]ustra0ivals'i | kstat'i ja8 t'ib'e0 [07]rasska0zɨval / [07]da0 / ka0k ja8 [09]e8ta / n:a [10]p'e0rvuju

orthographic [109]: [09a]во1т / э- / я1 та1к всегда1 и [13]устра1ивал | кстати я2 тебе1 [07]расска1зывал / [07]да1? / ка1к я2 [09]э2то / н:а [10]пе1рв

disfluencies [672]: э- | 9

technical_inform [69]

to avoid using them for further analysis. There were two types of defective speech fragments: segments of clipped signal and utterances with a maximum amplitude lower than one half the amplitude range. Such fragments were labelled as defective, but not excluded from the corpus.

3. Annotation

The monologues were left unannotated. In the read texts pauses were labelled automatically, with manual correction where necessary; additionally, when the speaker had problems with reading some words or phrases and re-read them, the defective fragment was segmented and labelled as a false-start.

The dialogues have the most detailed annotation. From each speaker's recording, 10–25 minutes of speech were selected for annotation. The participants' speech was annotated separately. The signal was manually segmented into chunks of up to a few seconds long. Ideally, a chunk should not contain defective signal (as defined in section 2.3.), long pauses, or partner's speech. However, due to very frequent clipping in some speakers' recordings, defective fragments shorter than 100 ms were allowed.

The signal between the chunks contains pauses, fragments of cross-talk, defective signal, or unannotated speech of the current speaker. Total duration values of all the segmented chunks for the given speaker are presented in table 1.

The annotation process was performed by five phoneticians using ELAN software (Wittenburg et al., 2006) (Hellwig, 2015) and took several stages:

1. orthographic annotation,
2. prosodic annotation,
3. segmental phonetic transcription,
4. boundaries of hesitations and non-speech events in audio signal.

This resulted in three tiers shown in Fig. 1.

Orthographic and prosodic annotation was performed manually (see 3.1. and 3.2. for more detail).

Segmental phonetic transcription—in accordance with the rules of Russian standard pronunciation (Avanesov, 1984)—was produced by automatic text transcriber with the orthographic transcription as an input. The transcriber has been developed at the Department of Phonetics, Saint-Petersburg State University, following the principles proposed by K. Shalonova (Shalonova, 1997).

Along with textual information, the orthographic tier contains hesitations and non-speech events. A new tier introduced manually provides information about their exact boundaries.

The annotation was checked by several passes of tests. Technical errors were assessed automatically including data format, misuse of special symbols, and spelling mistakes. Orthographic annotation was "peer-reviewed".

The annotation files are stored in ELAN-readable XML format and Praat TextGrid format (Boersma and Weenink, 2015).

3.1. Orthographic annotation

It is known that in real speech prosodic and syntactic unit boundaries do not necessarily coincide. It is not a sentence which is defined as the principal unit of speech, but rather an utterance or a phrase. Therefore the transcribers were asked to produce the orthographic transcription of the recording using no capital letters or punctuation marks;

the only exception was a question mark to denote question phrases.

Each word was written using standard spelling no matter whether it was pronounced in a proper way, mispronounced, or produced in a contracted form. Standard spelling was also used for commonly contracted forms, such as /grʲit/ instead of /ɡɑvɑˈrʲit/ for “говорит” (“says”), /t͡ɕek/ instead of /t͡ɕilɑˈvʲek/ for “человек” (“man”) etc.

Despite the ambiguous status of the grapheme “ё” in modern Russian, in this corpus it was never replaced by “е” (which is often the case in written texts). We consider it necessary since the two graphemes, “ё” and “е”, represent different phonemes.

If the word was completely unclear for the annotator, only the word’s rhythmic structure was written in CV-sequences.

Orthographic annotation also contained information about lexical stress: strong (primary) stress was marked with “1” after the vowel. Symbol “2” was used for vowels carrying secondary or weak stress and for unstressed vowels /o/, /e/ with no qualitative reduction.

Stress was marked according to the actual pronunciation, e. g. the word “звонит” (“calls”) could be transcribed as either “звони1т” or “зво1нит”, despite the fact that the latter variant is considered non-standard. The same is true for long words with a primary and a secondary stress. E. g., “среднеперсидский” (“Middle Persian”), typically pronounced with a primary stress on the second part (“сре2днеперси1дский”), could also appear with a primary stress on the first syllable (“сре1днеперси2дский”) to express contrast or emphasis on the first part. Furthermore, placing two strong stresses within one orthographic word was not forbidden, since it is quite possible in colloquial speech. Typical cases include hyphenated words, such as “ка1мень-на-оби1” (“Kamen-na-Obi”, proper name for a town in Russia) and “религио1зно-филосо1фский” (“relating to religion and philosophy”).

Various kinds of speech disfluencies were reflected in the orthographic transcription, e.g. elongations, hesitations, false-starts, self-corrections, non-speech events. All non-speech events were labelled with the symbol “9”. All hesitations were marked with the symbol “э-”, regardless of their actual pronunciation. Elongations were marked with a colon (:) after the lengthened sound. Self-corrections and false-starts were marked with an asterisk (*) at the end of the unfinished word or phrase. For words interrupted by a pause, hesitation or non-speech event, but finished after it, a caret (ˆ) was written at the end of the interrupted part, e.g. “восьмиˆ э- дне1вный” (“восьмидневный”, “eight-day”).

3.2. Prosodic annotation

Prosodic annotation was performed using a modified version of prosodic annotation system developed by N. B. Volskaya (Volskaya and Skrelin, 2009). The total number of prosodic models for all the dialogues in the corpus and a short description of models is given in table 2. A detailed description of prosodic annotation in CoRuSS can be found in (Volskaya and Kachkovskaia, 2016).

The system contains 13 basic contour types with up to four subtypes for each of them; the contours are described in both acoustic and pragmatic terms. The use of such a detailed classification for prosodic annotation enables us to analyse subtle individual differences and obtain a more detailed description of spontaneous speech intonation.

The speech string was divided into intonational phrases (IPs). Then for each IP the lexical word carrying nuclear accent was marked and the melodic type was assigned. This was performed using perceptual and acoustical data. Some IPs did not contain nuclear accent—typically, if the speaker failed to finish the IP.

The model type was placed in square brackets immediately before the word carrying nuclear stress. In hyphenated words with more than one stress, the model type was written immediately before the stressed component: e. g. “и она1 говори1т везё1м матра1сы в ка1мень-на-[01b]оби1” (and she says we’re carrying mattresses to Kamen-na-[01b]Obi).

If the intonational phrase contained more than one perceptually prominent word, such additional prominence was marked with the symbol [+].

E. g., the fragment “[+]о1чень у1мная [11]де1вочка / про1сто [10]вообще1 така2я / [+]блестя1щая [12]девчо1нка” (a [+]very smart [11]girl / in a [10]broad sense / [+]brilliant [12]girl) contains three IPs with a rise-fall in the first, a fall in the second, and a high rise in the third IP, with additional prominence in the first and the third IPs.

4. Statistics

The annotated part—subcorpus of dialogues—contains over 127,000 running words and over 83,000 accentual phrases, with the number of different word forms around 19,000. The total number of intonational phrases is around 45,000, of which almost 10,000 are entirely made of hesitations or non-speech events.

The frequency of hesitations and non-speech events (calculated relative to the number of orthographic words) is 4 % and 11 % respectively. Around 3 % of words contain elongations. Self-corrections are observed in 2 % of words.

In total, 5 % of all IPs did not contain nuclear stress. The percentage of IPs with additional prominence is 5 %. Average length of IP is 3.58 lexical words, or 2.31 prosodic words.

5. Conclusions

The corpus presented in this paper may serve a base for a wide range of empirical studies, including phonetic and prosodic inter- and intra-speaker variability, supervised learning for automatic speech recognition, spontaneous speech prosody, speech micro- and macro-planning strategies, relation between syntactic phrasing and dialogue acts. National licensing is currently in progress which will enable us to provide open access for other research groups within the field or beyond it.

6. Acknowledgment

This research was financed by a research grant 14-18-01352 “Automatic segmentation of speech signal into suprasegmental speech units” provided by Russian Scientific Foundation.

Table 2: Prosodic models: description and statistics.

Model	N	Description and *usage*
01	3	Very low fall; *signalling the end of a paragraph.*
01a	195	Low fall; *signalling the end of an utterance.*
01b	1462	A fall to non-low; *indicating cohesion, a link to the following utterance.*
02	4085	An intensified fall from a higher level; *used for giving emphasis.*
02b	141	Low-pitched stressed syllable and a rise-fall in the post-nuclear syllable; *used for giving extra emphasis.*
02c	1096	A fall from a high to mid or low level, accompanied by low intensity; *used to convey involvement and personal contact with the listener.*
03	86	Falling intonation for *wh-questions*, with the nuclear fall on the interrogative word.
03a	26	Falling intonation for *wh-questions*, with the nuclear fall not on the interrogative word but on some other word within the phrase.
04	180	Falling intonation with a wider interval of the falling tone, a higher level of intensity and appropriate voice quality (timbre); *used in exclamations.*
04a	35	Falling intonation with a wider interval of the falling tone, a higher level of intensity; *used in addressing a person.*
04b	41	Falling intonation with a wider interval of the falling tone, a higher level of intensity and special voice quality (timbre); *used in imperatives.*
05	35	A pitch rise to a very high level at some point in the pre-nuclear part, high plateau sustained up to the nucleus, a fall on the nucleus, normally accompanied by high intensity; *used in exclamations.*
06	5	Low-level; normally, with low intensity and increased vowel duration; *used in exclamations.*
06a	5	High-level; with low pre-nuclear part and increased vowel duration; *used in exclamations.*
06b	5	Mid-level; with low pre-nuclear part and increased vowel duration; *used in repeated or clarifying questions.*
06c	6	High-level; *used in emotional questions expressing disbelief or perplexity.*
07	466	A rise(-fall) on the last word in the IP; *used in general questions.*
07a	73	A rise(-fall) not on the last word in the IP; *used in general questions.*
07b	20	A rise with the F0 maximum shifted to the next syllable after nucleus; *used in general questions.*
08	73	Low (fall)-rise; *used in questions with an implied contrast.*
09	2028	(Low) level tone; *used for parentheses and author's remarks.*
09a	538	(Low) falling tone; *used for parentheses and author's remarks.*
09b	39	(Low) rising tone; *used for parentheses and author's remarks.*
10	4440	Non-low fall; *as, for example, at the end of a compound sentence component.*
11	10238	A rise(-fall), normally at a smaller interval compared to type 07; *used in non-final IPs.*
11a	436	A rise(-fall) with the F0 maximum shifted to the next syllable after nucleus; *used in non-final IPs.*
11b	1102	A rise(-fall) with a displaced F0 peak, often used for emphasis, but now commonly found in neutral speech of the younger generation; *used in non-final IPs.*
12	3329	A rise to a high pitch on the nuclear syllable levelled off in the post-nuclear part; *used in non-final IPs.*
12a	1864	A mid-level tone, actually realized as a step from a high pitch in the pre-nuclear part down to a medium pitch in the nucleus and sustained in the post-nuclear part; *used in non-final IPs.*
13	1912	Low (fall)-rise; *used in non-final IPs.*

7. Bibliographical References

Apresjan, J., Boguslavsky, I., Iomdin, B., Iomdin, L., Sannikov, A., and Sizov, V. (2006). A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects 1. In *Proceedings of LREC 2006. 5th International Conference on Language Resources and Evaluation*, pages 1378–1381.

Avanesov, R. (1984). *Russian Standard Pronunciation [Russkoe literaturnoe proiznoshenie]*. Prosveschenije.

Boersma, P. and Weenink, D. (2015). Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat.

Bogdanova-Beglarian, N., Martynenko, G., and Sherstinova, T. (2015). The "one day of speech" corpus: Phonetic and syntactic studies of everyday spoken russian. In *17th International Conference on Speech and Computer, SPECOM 2015*, volume 9319 of *LNAI*, pages 429–437. Springer.

Hellwig, B. (2015). Elan - linguistic annotator version

4.9.2. http://www.mpi.nl/corpus/html/elan/index.html.

A. A. Kibrik et al., editors. (2009). *Rasskazy o snovidenijakh. Korpusnoe issledovanie ustnogo russkogo diskursa.* Jazyki slavyanskoj kultury.

Krivnova, O. (2013). Russkij rechevoj korpus ruspeech. In *Proceedings of the VII International Scientific Conference "Fonetika segodnia"*, pages 54–56.

Shalonova, K. (1997). Flexible transcriber for russian continuous speech. In *2nd International Conference on Speech and Computer, SPECOM 1997*, pages 171–175.

Skrelin, P., Volskaya, N., Kocharov, D., Evgrafova, K., Glotova, O., and Evdokimova, V. (2010). Corpres – corpus of russian professionally read speech. In *13th International Conference Text, Speech and Dialogue, TSD 2010*, volume 6231 of *LNCS*, pages 392–399. Springer.

Volskaya, N. and Kachkovskaia, T. (2016). Prosodic annotation in the new corpus of Russian spontaneous speech CoRuSS. In *Proceedings of Speech Prosody 8, Boston, USA*.

Volskaya, N. B. and Skrelin, P. A. (2009). Prosodic model for russian. In *Proceedings of Nordic Prosody X*, pages 249–260. Peter Lager.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of LREC 2006. 5th International Conference on Language Resources and Evaluation*, pages 1156–1559.