Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6556–6561
Marseille, 11–16 May 2020

SibLing Corpus of Russian Dialogue Speech Designed for Research on Speech Entrainment

Tatiana Kachkovskaia, Tatiana Chukaeva, Vera Evdokimova, Pavel Kholiavin, Daniil Kocharov, Natalia Kriakina, Anna Mamushina, Alla Menshikova, Svetlana Zimina
Saint Petersburg State University
Saint Petersburg, Russia
kachkovskaia@phonetics.pu.ru, chukaeva68@gmail.com, postmaster@phonetics.pu.ru,
p.kholyavin@mail.ru, kocharov@phonetics.pu.ru, natalyakryakina@gmail.com,
mamushina.anna@mail.ru, menshikova.alla2016@yandex.ru, svetlanazimina6306@gmail.com

Abstract

The paper presents a new corpus of dialogue speech designed specifically for research in the field of speech entrainment. Given that the degree of accommodation may depend on a number of social factors, the corpus is designed to encompass 5 types of relations between the interlocutors: those between siblings, close friends, strangers of the same gender, strangers of the other gender, strangers of which one has a higher job position and greater age. Another critical decision taken in this corpus is that in all these social settings one speaker is kept the same. This allows us to trace the changes in his/her speech depending on the interlocutor. The basic set of speakers consists of 10 pairs of same–gender siblings (including 4 pairs of identical twins) aged 23–40, and each of them was recorded in the 5 settings mentioned above. In total we obtained 90 dialogues of 25–60 minutes each. The speakers played a card game and a map game; they were recorded in a soundproof studio without being able to see each other due to a non-transparent screen between them. The corpus contains orthographic, phonetic and prosodic annotation and is segmented into turns and inter-pausal units.

Keywords: speech corpus, speech entrainment, phonetics, dialogue, siblings, twins

1. Introduction

In recent years, there has been an increasing interest in speech entrainment research. The studies on the topic contribute to understanding of the processes at work in human–human communication and offer perspectives for the imitation of these processes in human–computer interaction. However, the problem of speech material for investigations in this field is still present.

The corpora recorded specifically for studies on entrainment include those for national varieties of English (Pardo, 2006) (Branigan et al., 2000) (Levitan et al., 2016) (Nenkova et al., 2008), German (Schweitzer and Lewandowski, 2013), French (Bailly and Lelong, 2010), Italian (Savino et al., 2016), Japanese (Kawahara et al., 2015) and other languages; Russian, however, is not one of these.

In the studies on entrainment, one of the most popular ways of arranging a conversation between participants is a task-oriented dialogue in a game form. The tasks are sometimes restricted in terms of spontaneity (Bailly and Lelong, 2010), but in most cases they encourage free communication (De Looze et al., 2014). The types of games include map tasks (Pardo, 2006), ranking objects important for survival in a dangerous situation (Kousidis et al., 2009), card matching (Szekely et al., 2015), matching Tangram figures (Savino et al., 2016). This form of interaction requires engagement and cooperation of participants which is expected to be a favourable condition for entrainment. Other ways of obtaining spontaneous speech for studying entrainment are conversations on free topics (Natale, 1975) or given ones. In the majority of cases speakers have no visual contact.

Another approach to material collection in this field is shadowing tasks (Babel, 2009) (Bulatov, 2009) (Pardo et al., 2013). One of the widespread types of this task requires a participant to repeat words after the model talker. The method allows for a more direct observation of entrainment; however, it lacks spontaneity.

The number of speakers in the existing corpora varies greatly. The largest groups of subjects are found in the experiments with shadowing tasks (e.g. in (Babel, 2009)—2 male model talkers, 117 shadowers (53 male, 64 female); in (Pardo et al., 2017)—12 model talkers (6 male, 6 female), 96 shadowers (48 male, 48 female)). As for dialogues, both task-oriented and free conversations, the number of participants ranges from about 10 to 25 (however, not all the data is balanced in terms of gender). Some authors report that the interlocutors were strangers to each other (Pardo, 2006) or were acquainted (Savino et al., 2016); rare experimental designs use the degree of familiarity as a variable (Schweitzer and Lewandowski, 2013).

The corpora differ in terms of annotation. Some are provided with orthographic transcriptions (Schweitzer and Lewandowski, 2013), prosodic annotation, segmentation on different levels (turn-level (Kawahara et al., 2015), speech–pause intervals (Savino et al., 2016), syllable and word levels (Schweitzer and Lewandowski, 2013).

It should be mentioned that many studies rely on the corpora originally designed for other purposes. The employed methods are similar to those described above and include dialogues—theme-based or free conversations and performing a broad variety of game tasks—e.g., (Reichel and Cole, 2016), (Levitan et al., 2015), (Cabarrão et al., 2016), (Karpiński et al., 2014) and etc. An apparent advantage of these corpora is their wide annotations. It is important to note that some of the corpora vary the de-

gree of familiarity between participants (e.g. in (Cabarrão et al., 2016)—from strangers to twin sisters, in (Ireland et al., 2011)—strangers and couples engaged in romantic relationship).

The corpus presented in this study is the first corpus of Russian speech recorded specifically for the purposes of entrainment research. This corpus, named SibLing, was designed taking into account the latest experience obtained by other research groups in the field. SibLing contains 90 dialogues from 100 speakers. Apart from the size of the corpus, one of its considerable advantages is that the data is balanced in terms of speakers' gender and degree of familiarity (strangers, friends, siblings, twins). The dialogues are in the form of games, namely, map tasks and card matching (finding two elements in common). The choice of this technique enables future comparison with the results obtained on other corpora using the same tasks.

The entire corpus has both orthographic and phonetic transcription; the annotation levels include turn-boundaries, inter–pausal units, keywords, and partially prosodic annotation. The SibLing corpus will be freely available to scientific groups. Please contact the first author to request access to the corpus.

2. Corpus design

The basic set of the dialogues was recorded from 10 pairs of same–gender siblings aged between 23[1] and 40 (5 male pairs, 5 female pairs); of these, there were 4 pairs of identical (monozygotic) twins (2 male pairs, 2 female pairs). The basic set of speakers is presented in Table 1.

Each of these speakers communicated with the following interlocutors:

1. their sibling, same gender, (approx.) same age;
2. a close friend, same gender, approx. same age;
3. a stranger, same gender, approx. same age;
4. a stranger, other gender, approx. same age;
5. a stranger of older age having a high job position, same gender.

All the interlocutors were native Russian speakers that now reside in Saint Petersburg; none of the speakers reported on any speaking or hearing impairment. In all pairs of interlocutors except for those from the group 5 the age difference did not exceed 4 years.

In total, the corpus comprises 90 dialogues. The recordings were obtained in the recording studio in a WAV format with 24-bit, 44100-Hz sampling frequency to ensure the quality of the material. Each speaker was recorded using AKG HSC 271, an individual headset equipped with a condenser microphone. Additionally, a bi-directional microphone (Audio-Technica AT 2050) was placed between the speakers. Thus, speech was recorded from three sources in multi-channel mode; the recorded speech was exported into three separate audio files. The interlocutors were separated by a non-transparent screen used to prevent them from seeing each other and the cards/maps of the other speaker.

[1] In Russia, 23 is the age when a person usually graduates from a university or college.

Speakers	Gender	Relation	Ages
S01, S02	F	twins	26, 26
S03, S04	F	siblings	36, 38
S05, S06	M	siblings	36, 38
S07, S08	M	siblings	31, 36
S09, S10	F	twins	38, 38
S11, S12	M	twins	24, 24
S13, S14	F	siblings	28, 30
S15, S16	M	siblings	33, 30
S17, S18	M	twins	23, 23
S19, S20	F	siblings	32, 33

Table 1: The basic set of speakers

Speakers completed two speaking tasks which lasted between 25 and 60 minutes in total. Task one was a **card game** based on searching for similarities in two decks of ten cards. The cards were very much like the ones from the famous Dixit card game designed by Jean-Louis Roubira, i.e. each card depicted several different objects combined in one dreamlike picture. One of the participants was instructed to describe his/her upper card to the interlocutor, while the other's task was to look through his/her deck and find a card with at least two matching objects. The decks were arranged randomly, yet had a number of similar objects depicted. The speakers took turns to describe their picture, thus swapping the Leader and the Follower roles several times. The game lasted for 10–12 minutes; then the speakers were asked to proceed to task two.

Task two was a **map task** in which the interlocutors were asked to guide each other through a set of schematic maps. Each pair of speakers was given four pairs of maps so that each speaker got two "complete" and two "incomplete" ones. A complete map had a route marked on it, whereas an incomplete one lacked the route; additionally, some landmarks differed or were replaced—see example on Figure 1. A person having a complete map was asked to explain the route to his/her interlocutor, and the latter had to draw it on his/her own incomplete map. After the explanation, the other person was instructed to repeat the route he/she had drawn. The speakers took turns to describe their complete map, thus swapping the Leader and the Follower roles 4 times.

Each pair of maps contained 5 keywords that were kept the same in all the recordings. All the keywords mostly contained vowels and sonorants (and very rarely voiced fricatives) to enable reliable comparison of prosody during further analysis. All the keywords are listed in Table 2.

As each speaker of the basic set of siblings participated in 5 recordings, each map set had 5 varieties, with keywords kept the same but the route and other landmarks differing.

After completing the tasks, speakers were asked to fill in two questionnaires. The first one was filled by each speaker only once and included personal questions: his/her age, gender, level of education, profession, birthplace, native

Map	Keywords
1	ювели́рный /juvʲiˈlʲirnɨj/ jeweller's, вое́нный райо́н /vaˈjennɨj raˈjon/ military district, мыловáрня /mɨlaˈvarnʲa/ soap factory, авиаýзел /avʲiaˈuzʲil/ air hub, мавзоле́й /mavzaˈlʲej/ shrine
2	равни́на /ravˈnʲina/ flatland, мали́новая алле́я /maˈlʲinavaja aˈlʲlʲeja/ raspberry lane, земля́ оле́ней /zʲimˈlʲa aˈlʲenʲij/ deer land, руи́ны /ruˈinɨ/ ruins, взмо́рье /ˈvzmorʲji/ seashore
3	муравéйник /muraˈvʲejnʲik/ ant hill, воро́ний волье́р /vaˈronʲij vaˈlʲjer/ crows' aviary, ви́нные зе́мли /ˈvʲinnɨji ˈzʲemlʲi/ wine lands, заозе́рье /zaaˈzʲerʲji/ across the lake, о́зеро Я́мное /ˈozʲira ˈjamnaji/ lake Yamnoye
4	ламина́рия /lamʲiˈnarʲija/ laminaria, валериа́на /valʲiˈrʲjana/ valeriana, ма́льва /ˈmalʲva/ malva, лимо́н /lʲiˈmon/ lemon tree, ало́э /aˈloe/ aloe

Table 2: Keywords for the map task

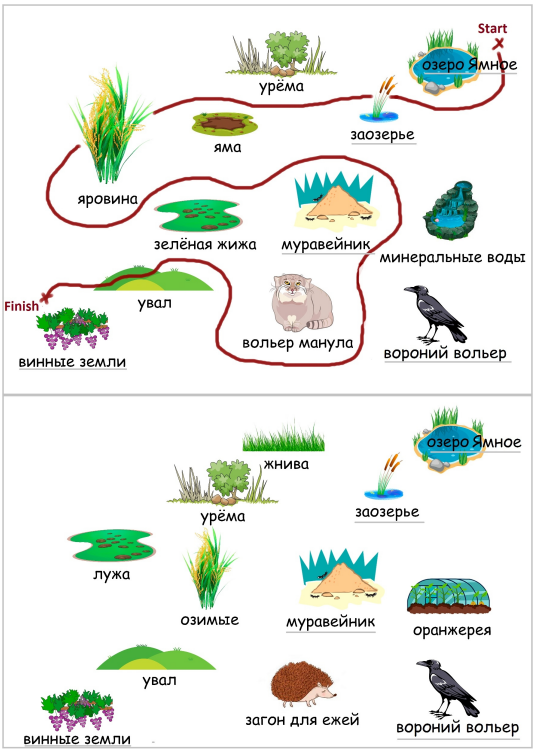

Figure 1: A pair of maps used in sibling–sibling recordings with keyword set #3 (see Table 2.) (Keywords are underlined here, but were not underlined on the prints used during the recording.)

language, experience in practicing pronunciation of foreign languages, cities where he/she spent their childhood/went to school/attended a university or college, and cities where he/she had lived for at least one year. The second questionnaire was filled after every recorded dialogue and contained a series of open questions intended to find out whether the speaker felt comfortable during this particular session and whether the task was completed successfully.

3. Annotation

The annotation scheme includes seven tiers:

1. clipped speech fragments
2. inter-pausal units
3. dialogue turns
4. fundamental frequency pulses
5. orthographic and prosodic transcription
6. phonetic transcription
7. keywords

Fragments of clipped speech were detected automatically based on speech signal amplitude with the threshold of 99.9 % of the maximum possible amplitude. A 20 ms long speech frame was defined as clipped if it contained more than two clipped amplitude peaks.

Inter-pausal units (IPUs) were detected automatically by means of the speech detection module provided within Kaldi ASR toolkit (Ghahremani et al., 2014). The small amount of error decisions produced by the automatic procedure was corrected manually.

Dialogue turns were detected automatically based on IPU boundaries. The IPU of a certain speaker was detected as the first IPU of a dialogue turn if it was pronounced within a pause in the interlocutor's speech. The main rule was that there could not be two consecutive turns pronounced by the same speaker. Backchannel turns were the only exception for this rule and were annotated with a special label. All the turn collisions that occurred due to overlap of turns and backchannels were corrected manually.

Fundamental frequency (F_0) within IPUs was calculated by means of the pitch detection module from Kaldi ASR toolkit. Of all the F_0 values calculated for each processing frame, we keep only those with the probability of voicing above 0.85. With the threshold this high, we aim to get rid of those pitch pulses that occur at segment boundaries and usually cause microprosodic events. The subsequent approximation of F_0 pulses was produced by means of linear interpolation and alignment of pulses along the F_0 contour of the voiced fragments. Voiceless fragments were bridged by means of linear interpolation. Then each F_0 pulse was labeled with the value of its fundamental frequency.

The orthographic and prosodic transcription of the IPUs was produced manually by professional linguists according to the rules developed earlier by our research team for another speech corpus (Kachkovskaia et al., 2016). The orthographic transcription was produced for all the dialogues. Punctuation marks were not used except for the question mark, apostrophe and hyphen. The spelling of a

Corpus	MTLD	Vocabulary	N of words
SibLing (all)	48.84	7169	52541
SibLing (cards)	59.93	4106	18636
SibLing (maps)	44.04	3793	33091
CoRuSS (part)	65.53	10765	52541
CoRuSS (all)	65.46	19231	119859

Table 3: Comparison of MTLD, size of vocabulary and number of words in CoRuSS and SibLing corpora. For SibLing, we also present data on card game only and map task only. CoRuSS (part) is a part with the same amount of running words as in the whole SibLing corpus.

word included stress indication according to standard pronunciation rules. Various speech disfluencies (including false-starts, self-corrections, word abruptions, hesitations), non-speech events (including laughter, meaningful clicks, coughs) were indicated as well.

Non-standard (occasional) pronunciations (e.g., with sound replacement—such as "жижда" instead of "жижа", slush) were indicated and added to a special list to enable access to both the pronounced variant and the intended one.

Then, prosodic annotation was added to the orthographic transcription. Using a special set of symbols, the annotators marked the intonational phrase boundaries, the nuclear accent and the type of the melodic movement within the nucleus. The prosodic annotation scheme is described in details in (Volskaya and Kachkovskaia, 2016) and is roughly in line with the British School's annotation principles (such as in the description provided by O'Connor and Arnold): an intonational phrase usually contains one nucleus, and the melodic movement within the nucleus is one of a set of values defined for this particular language. As manual prosodic annotation is a very time consuming procedure we annotated only the card game dialogues (approx. 12 minutes for each pair of speakers).

Broad phonetic transcription—in accordance with the rules of Russian standard pronunciation (Avanesov, 1984)—was produced by the automatic text transcriber with the orthographic transcription as an input. The transcriber used here was developed at the Department of Phonetics, Saint-Petersburg State University (Evdokimova et al., 2017).

The keyword tier includes manual annotation of physical boundaries of the keywords. The keyword list was defined a priori—see Table 2.

4. General overview of the data

4.1. Lexical diversity

We estimated lexical diversity with the measure of textual lexical diversity (MTLD) proposed by McCarthy (McCarthy and Jarvis, 2010). This measure allows comparing text of various lengths, whereas simple Text-Type Ratio (TTR) has a strong negative correlation with text length. As an input we used the orthographical transcription of the dialogues lemmatized by Pymorphy2 (Korobov, 2015). Table 3 presents lexical diversity of the recorded dialogues in comparison with the Corpus of Russian spontaneous speech (CoRuSS) (Kachkovskaia et al., 2016). CoRuSS is an annotated collection of free-topic dialogues by sixty speakers.

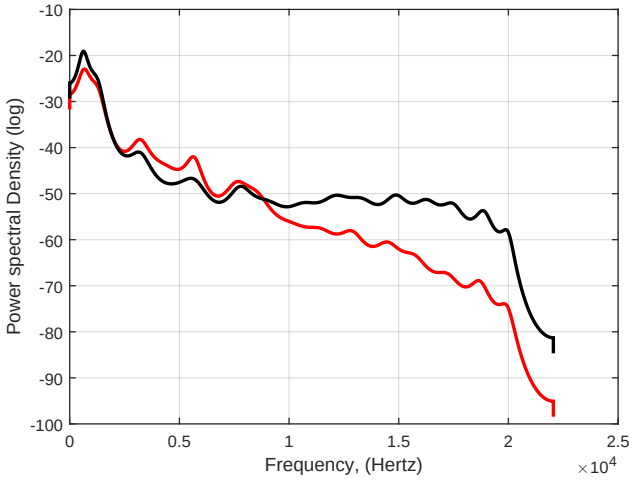

Figure 2: Long-term average spectra for a pair of female twins: S01 and S02.

The main observation is that lexical diversity within the SibLing corpus is smaller than the one of CoRuSS data. It is an expected result as here dialogues are limited in topics [2]. This becomes particularly clear when we compare data calculated separately for the card game and for the map task. With the card game, the lexical diversity is closer to the value obtained for CoRuSS which contains free–conversation dialogues. With the map task, on the other hand, the lexicon is very limited—first, due to object names directly written on maps, and second, due to multiple repetitions of direction words ("right", "left", "turn", "go around" etc.). This observation enables to use different parts of the SibLing corpus for specific tasks. Thus, the lexically poor map task subcorpus makes it easy to look for signs of speech entrainment in the lexicon (choosing between synonyms) and compare pronunciations of the same words by different speakers.

4.2. Speaker difference and similarity

To roughly analyze the differences and similarities between the speakers, we calculated long-term average spectra (Löfqist and Mandersson, 1987) for female twins (Figure 2), female siblings (Figure 3) and female non-relative speakers (Figure 4). The correlation analysis of the plotted spectra resulted in following:

- twins (S01 and S02): r=0.98
- same-sex siblings (S03 and S04): r=0.96
- non-related speakers (S03 and S05): r=0.91
- non-related speakers (S03 and S08): r=0.92
- non-related speakers (S05 and S08): r=0.91

The preliminary results of correlation analysis show that same-sex close-age siblings speak almost with the same similarity of the voice as twins do. The corpus will make it possible to analyze entrainment of both twins and close-age siblings speaking in similar conditions.

[2]For linguistic properties, including lexical diversity, of different types of conversational tasks see (Pallotti, 2019)

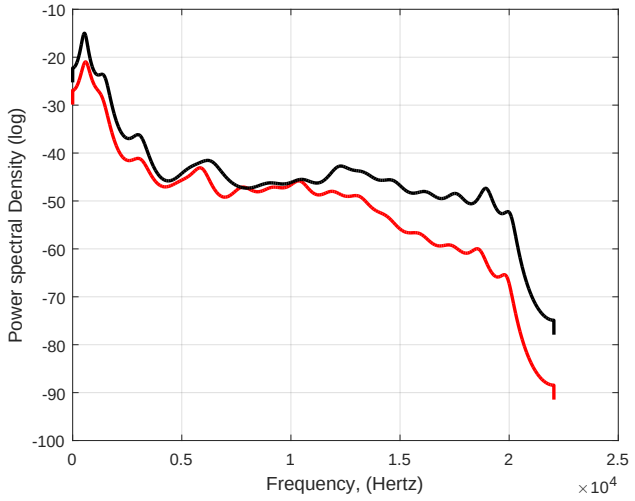

Figure 3: Long-term average spectra for a pair of female siblings S03 and S04.

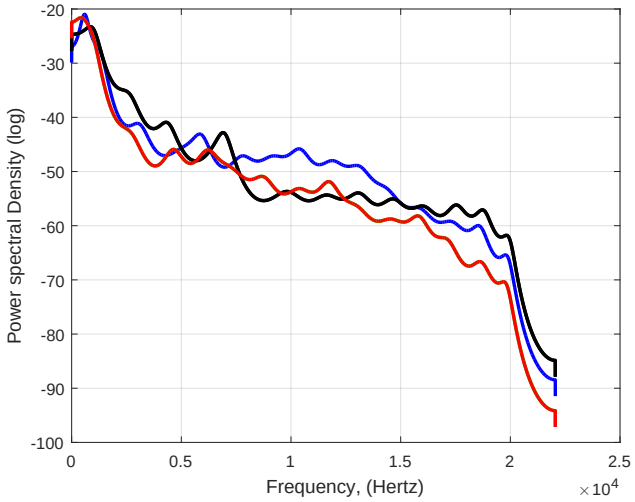

Figure 4: Long-term average spectra for three female non-relative speakers: S03, S05, S08.

5. Conclusion

The corpus presented in this paper, SibLing, is specifically designed to solve various tasks in the field of speech entrainment. Based on the up-to-date publications, the corpus includes dialogues between speakers of 5 degrees of "closeness" (familiarity)—from siblings to strangers of different social status. This design, along with the impressive size of the corpus and balance in gender, makes SibLing a unique resource.

Being a spontaneous speech corpus, SibLing is well-suited for solving a broad variety of other tasks within the fields of both phonetics and speech technology. Phonetics tasks include investigating various phenomena of spontaneous speech, such as disfluencies and strategies of turn-taking. Of particular interest is research on how a speaker's voice changes in different social situations—in our case, depending on who he/she is speaking with; knowing the limits of variability for one speaker are crucial in solving speaker verification tasks. As for speech technologies, apart from entrainment research, the SibLing corpus will be useful for a range of tasks in the field of speaker verification.

6. Acknowledgements

The research is supported by the Russian Science Foundation (grant 19-78-10046 "Phonetic manifestations of communication accommodation in dialogues").

7. Bibliographical References

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

Babel, M. (2009). Selective vowel imitation in spontaneous phonetic accommodation. In *UC Berkeley Phonology Lab Annual Report 2009*, pages 163–194.

Bailly, G. and Lelong, A. (2010). Speech dominoes and phonetic convergence. In *Proceedings of Interspeech 2010*, pages 1153–1156.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75(2):13–25.

Bulatov, D. (2009). The effect of fundamental frequency on phonetic convergence. In *UC Berkeley Phonology Lab Annual Report 2009*, pages 404–434.

Cabarrão, V., Trancoso, I., Mata, A., Moniz, H., and Batista, F. (2016). Global analysis of entrainment in dialogues. In *Proceedings of IberSPEECH 2016*, pages 215–223.

De Looze, C., Scherer, S., Vaughan, B., and Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58:11–34.

Evdokimova, V., Chukaeva, T., and Skrelin, P. (2017). Automatic phonetic transcription for Russian: Speech variability modeling. In *Proceedings of SPECOM 2017*, pages 192–199.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., and Khudanpur, S. (2014). A pitch extraction algorithm tuned for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2494–2498.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., E., S. L., Finkel, E. J., and Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

Kachkovskaia, T., Kocharov, D., Skrelin, P., and Nina, V. (2016). CoRuSS—a new prosodically annotated corpus of Russian spontaneous speech. In *Proceedings of LREC 2016*, pages 1949–1954.

Karpiński, M., Klessa, K., and Czoska, A. (2014). Local and global convergence in the temporal domain in Polish task-oriented dialogue. In *Proceedings of the 7th Speech Prosody Conference*, pages 743–747.

Kawahara, T., Yamaguchi, T., Uesato, M., Yoshino, K., and Takanashi, K. (2015). Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening. In *Proceedings of APSIPA Annual Summit and Conference 2015*, pages 392–395.

Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In *Analysis of Images, Social Networks and Texts*, pages 320–332.

Kousidis, S., Dorran, D., McDonnell, C., and Coyle, E. (2009). Times series analysis of acoustic feature convergence in human dialogues. In *Proceedings of SPECOM 2009*, pages 1–6.

Levitan, R., Beňuš, S., Gravano, A., and Hirschberg, J. (2015). Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison. In *Proceedings of the SIGDIAL 2015 Conference*, pages 325–334.

Levitan, R., Beňuš, S., Gálvez, R., Gravano, A., Savoretti, F., Trnka, M., Weise, A., and Hirschberg, J. (2016). Implementing acoustic-prosodic entrainment in a conversational avatar. In *Proceedings of Interspeech 2016*, pages 1166–1170.

Löfqist, A. and Mandersson, B. (1987). Long-time average spectrum of speech and voice analysis. *Folia Phoniatr*, 39:221–229.

McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790–804.

Nenkova, A., Gravano, A., and Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 169–172.

Pallotti, G. (2019). An approach to assessing the linguistic difficulty of tasks. *Journal of the European Second Language Association*, 3(1):58–70.

Pardo, J., Duran, K., Mallari, R., Scanlon, C., and Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, 69:183–195.

Pardo, J. S., Urmanche, A., Wilman, S., and Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention Perception and Psychophysics*, 79(2):637–659.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Association of America*, 119(4):2382–2393.

Reichel, U. and Cole, J. (2016). Entrainment analysis of categorical intonation representations. In *Proceedings of Phonetik and Phonologie*, pages 165–168.

Savino, M., Lapertosa, L., Caffò, A., and Refice, M. (2016). Exploring prosodic convergence in Italian game dialogues. In *Proceedings of the 7th Tutorial and Research Worshop on Experimental Linguistics (Exling 2016)*, pages 151–154.

Schweitzer, A. and Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. In *Proceedings of Interspeech 2013*, pages 525–529.

Szekely, E., Keane, M., and Carson-Berndsen, J. (2015). The effect of soft, modal and loud voice levels on entrainment in noisy conditions. In *Proceedings of Interspeech 2015*, pages 150–155.

Volskaya, N. and Kachkovskaia, T. (2016). Prosodic annotation in the new corpus of Russian spontaneous speech CoRuSS. In *Proceedings of Speech Prosody 2016*, pages 917–921.