The "One Day of Speech" Corpus: Phonetic and Syntactic Studies of Everyday Spoken Russian

Natalia Bogdanova-Beglarian, Gregory Martynenko, and Tatiana Sherstinova(✉)

St. Petersburg State University,
Universitetskaya Nab. 11, St. Petersburg 199034, Russia
nvbogdanova_2005@mail.ru, {g.martynenko,sherstinova}@gmail.com

Abstract. The studies described in the paper are made on the base of the ORD – "One day of speech" – corpus of Russian everyday speech which contains long-term audio recordings of daily communication. The ORD corpus provides rich authentic material for research in phonetics and syntax of spoken Russian, and may be used for adjustment and improvement of speech synthesis and recognition systems. Current phonetic investigations of the ORD corpus relate to temporal studies, study of speech reduction, phonetic realization of words and affixes, investigation of phonetic errors and mondegreens, studies of rhythm structures and hesitation phenomena. Syntactic studies primarily deal with linear word order of syntactic groups, syntactic complexity of spoken utterances, and specific syntactic phenomena of spontaneous speech. In this paper, we summarize main achievements in phonetic and syntactic studies made on the base of the ORD corpus and outline some directions for further investigations.

Keywords: Everyday spoken Russian · Speech corpus · Phonetics · Syntax

1 Introduction: The ORD Corpus

Everyday spoken Russian has been the subject of scientific analysis since the works of E. Zemskaja, O. Sirotinina, O. Lapteva, N. Rozanova, M. Kitajgorodskaja, and other linguists. However, up to present, there were not enough linguistic resources of Russian real-life spontaneous speech. For example, the Spoken Speech Subcorpus in Russian National Corpus does not contain any audio data at all, consisting just of speech transcripts [1]. The other well-known Night Dream Stories corpus contains both texts and thoroughly annotated speech recordings. However, this corpus is relatively small (about 2 h of recordings, 14000 words in transcripts) and contains a restricted number of spoken genres (mainly narratives) [2]. There are also other Russian speech corpora that should be mentioned: RuSpeech corpus [3], an annotated corpus of Russian

A. Ronzhin et al. (Eds.): SPECOM 2015, LNAI 9319, pp. 429–437, 2015.
DOI: 10.1007/978-3-319-23132-7_53

speech [4], corpus of emotion Russian speech [5] and some other resources. However, all these corpora either contain no everyday speech recordings or are limited just to few communicative situations (like the latter one).

The only Russian corpus containing spoken everyday speech recorded in natural and diverse communicative situations is the "One Day of Speech" (ORD) corpus that has its origin in St. Petersburg State University. The recordings are made by participants-volunteers who spend a whole day with switched-on voice recorders and record all their audible communication [6]. The similar methodology of long-term recordings had been earlier used for collecting data for the British National Corpus [7] and the JST ESP corpus in Japan [8].

Speech is transcribed and selectively annotated on different levels – phonetic, lexical, grammatical, and pragmatic levels. Transcribing and most annotations are made in ELAN [9]. Phonetic annotation is made in Praat [10]. Quantitative data processing is made for annotations on each level [11].

Nowadays, the ORD corpus is one of the most representative collections of everyday spoken Russian. It contains more than 1000 h of recordings gathered from 110 main participants and hundreds of their interlocutors. The ORD volunteer participants represent various professional and status strata. The age of participants rangers from 18 to 77 years with an average value of 37 years. Speech transcripts comprise about 500000 words.

The ORD corpus provides rich authentic material for research in phonetics and syntax of spoken Russian, and for solving applied linguistic problems in speech technologies (e.g., it may be used for adjustment and improvement of speech synthesis and recognition systems, and for forensic phonetics). In this paper we summarize main achievements in phonetic and syntactic studies made on the base of this corpus and outline some directions for further investigations.

2 Phonetic Studies

The initial goal of the ORD corpus was to conduct phonetic studies of Russian everyday speech. Nowadays, we may list the following phonetic aspects that are being successively investigated on the ORD data: temporal studies, study of speech reduction, phonetic realization of words and affixes, investigation of phonetic errors in speech production and mondegreens in speech perception, studies of rhythm structures and hesitation phenomena.

2.1 Temporal Studies

All multimedia annotations of speech signal are made in linguistic annotator ELAN, therefore each annotated phenomenon (sound, morpheme, word, phrase, turn, etc.) refers to a particular segment in correspondent sound file and has particular duration. Thus, temporal study of elements is possible on all linguistic levels. For example, on phonetic level we study speech rate, rhythmic patterns, temporal registers of Russian everyday speech, and other temporal phenomena.

In our phonetic studies, first of all, we obtained the frequency distribution of utterance length in words. Based on the ORD data, the average utterance length for spoken Russian is 4.35 words (SD = 4.02). Most of all utterances consist of a single word or a word-like particle (25.26 %). Two-word utterances make 15.58 % of the whole data, three-word utterances have the third rank that makes 12.45 %. Four-word utterances make 10,98 %, five-word utterances – 8.74 %, etc. [12].

If to measure utterance length in syllables, it turns out that the most frequent Russian utterances consist of one or two syllables and represent 11.0 % and 11.7 % of all utterances respectively. Three-syllable utterances are ranked third (8.7 %), four-syllable utterances take 7.7 %, five-syllable – 6.88 % and six-syllable utterances – 6.50 %. Utterances longer than 20 or more syllables take up less than 1 % of the whole. Thus, more than half of all spoken communication consists of short utterances with a length up to 6 syllables [12].

Therefore, the majority of Russian spoken interaction consists of one or a few word utterances that contain one or few syllables. As for the dependency of average utterance duration on their length in syllables, it is well described by the following linear function: $y = 133.28x + 367.5$, where x is a number of syllables, and y is an average duration of such utterances in milliseconds [12].

An average tempo of ORD informants is 5.31 syllables per second (syl/s). The variation among informants begins from the slowest 3.6 syl/s till the fastest 6.7 syl/s. These numbers are on average higher than, for example, in Norwegian (3.5–4.5 syl/s), in standard northern Dutch (5.2 syl/s), or in French (according to some data, 4.31 syl/s). However, they are significantly lower than in Spanish (7.81 syl/s) or in Brazilian Portuguese (6.57 syl/s). As Russians in Russia, with approximately the same tempo, people speak English in the UK (3.16–5.33 syl/s), as well as in the USA (3.1–5.4 syl/s) [13].

It was determined that there are several factors that influence speech tempo in Russian: gender (men speak faster than women), age (the older a speaker, the slower he or she speaks), level of language competence (the higher the competence level, the slower is the speech), and social role of speakers (speech is faster when communicating with friends than in work settings) [13,14].

Finally, the hypothesis on existence of two temporal registers of speech was proposed: (1) the "regular" (or dialogue) register is used for producing utterances whose length does not exceed 15 syllables. Its distinctive feature is a strong interrelation between an average syllable duration (syllable rate) of utterances and their length in syllable (in this case the average syllable duration is a function from utterance length in syllables, ranging from 450 ms to 150 ms) and (2) the "speedy" (or monologue) register is used for producing longer utterances (exceeding 15 syllables). In contrast to a dialogue register, the average syllable duration (or an average utterance rate) of a "speedy" register does not depend on utterance length in syllables and is equal to approximately 150 ms [12].

2.2 Study of Reduction. Phonetic Realization of Words and Affixes

Reduced forms of different words and phrases, especially the most commonly used in everyday Russian, are analyzed. Many of these forms have already got

correspondent written forms, which may be found in modern Russian literature and which are frequent in electronic communication. Current analysis touched on all possible features of reduced forms: (1) pronunciation (i.e., detailed phonetic transcription), (2) spelling (common variants of written forms), (3) semantic features, and (4) pragmatic features. Based on the results, the multimedia dictionary of reduced forms in Russian is created [15]. The study of spontaneous speech reduction may be used for building an authentic lexicon of word pronunciations.

In search for the correlation between grammatical meanings of morphemes and their phonetic realization, the real phonetic transcription of inflectional affixes for different speakers in various communication situations was obtained [16]. Statistical tables are drawn with correspondence of orthography, real phonetic transcription, and grammatical categories. These data provided the basis for the Audio dictionary of Russian inflectional affixes.

2.3 Studying the "Weak Points" in Speech Perception and Production

The lists of common mistakes of hearing (mondegreens or incorrect attribution of words) that were made by linguists-experts while transcribing the ORD recordings were compiled [17]. Based on these lists, the analysis of significant perceptive elements of word forms that are essential for their proper attribution was made. The following elements turned out to be invariant despite incorrect recognition of words: stressed syllable position, segmental parts of the stressed syllable (both consonants and vowels), number of syllables, consonantal "skeleton" of word form, its either initial or final segmental fragment.

The analysis of the phonetic mistakes that are typically done by Russian speaking people in everyday speech was made, too. Most common mistakes are the following: incorrect stress position in words and phrases, the alignment of phonetic features in neighboring words, palatalization errors, substitution of proper words by not-existing quasi-words, which are phonetically similar to the prototype words. However, we should point out that phonetic mistakes occur comparatively less than other type of errors, which have been also analyzed. Cf.: phonetic errors – 17 %, lexical errors – 28 %, morphology errors – 31 %, and syntactic errors – 24 % [18].

2.4 Russian Speech Rhythm Studies

Empirical investigation of the ORD recordings has revealed a tendency towards the usage of symmetrical rhythmical structures built of isochronic or quasi-isochronic segments in Russian everyday speech [19]. The most typical are structures consisting of two, three, four, etc. quasi-isochronic segments. Moreover, it was observed that phenomenon of isochronism of speech rhythmic structures can simultaneously appear on different structural levels. We have a hypothesis that the lower quasi-isochronic level performs the role of some kind of inner metronome, which organizes our speech flow. It may change its tempo on borders of rhythmic groups. However, these structural levels do not correlate with

linguistic levels. We may suggest that the distribution of linguistic units onto "isochronic boxes" is determined mainly by pragmatics: the more important the segment is the more "boxes" it may take [19].

The use of fillers and other discourse markers in spontaneous speech may be explained in many cases just by unconscious desire of speakers to reach this temporal pattern. The examples of such cases are given and explained in [20]. This hypothesis is to be tested on representative ORD data.

2.5 Hesitation Phenomena

Speech hesitations are a common feature of spontaneous speech production. According to the ORD data, both filled and silent hesitation pauses are among the most frequent elements of spoken Russian. They naturally occur in all types of speech and by all speakers. Different types of hesitations found in the corpus have been analyzed. The classification model of hesitation phenomena was proposed. The most frequent hesitations are the following: silent pauses, stretching of sounds, interruptions, repetitions, filler-words, other kinds of fillers, and paralinguistic actions. The new term "verbal hesitation" has been introduced for denotation of verbal fillers of hesitation nature [21].

3 Syntactic Studies

Syntax is a part of grammar where the features of spoken language are most clearly revealed in a variety of ways (for example, cf. [22]). Syntactic studies of spontaneous speech are very important for ASR systems. Nowadays, language models of the most speech recognition systems are trained on the corpora of written texts. However, written Russian and spoken Russian differ greatly from each other in respect to some fundamental syntactic features, as it is shown below. Therefore, n-gram models that are built for written language cannot be efficiently applied to LVCSR tasks. That particularly refers to recognition of spontaneous real-life speech. Speech transcriptions of the ORD corpus form a valuable resource for creating a language model of spoken Russian.

Several syntactic studies have been already made on the base of the ORD corpus. One of them is a pilot research of verbal groups in Russian spontaneous speech. Based on the random sample of 550 verbal branches represented in the formal way, the following models of left- and right-branching subordinations were found:

1. Verbs without dependents (*V*): 13.64 %;
2. Symmetrical verbal groups (*1V1, 2V2, 3V3*): 14.00 %;
3. Generally left-branching verbal groups (*1V, 2V, 3V, 4V, 5V, 2V1, 3V1, 3V2, 4V1*): 59.27 %;
4. Pure left-branching verbal groups (*1V, 2V, 3V, 4V, 5V*): 50.36 %;
5. Generally right-branching verbal groups (*V1, V2, V3, V4, V5, 1V2, 1V3, 1V4, 2V3*): 12.90 %;

6. Pure right-branching verbal groups (*V1, V2, V3, V4, V5*): 7.45 %.

We have calculated the averages, characterizing left- and right- branching in verbal groups. Thus, the average left-width of the branch equals to 1.195, the average right-width of the branch is 1.565. Their ratio (L/R) equals to 1.309, therefore we observe in spoken Russian an evident trend towards left-branching asymmetry.

These results are very different from data obtained on the material of written texts. For example, in [23] is shown that in written Russian texts the ratio of left-branching structures to right-branching ones is close to 1 (i.e., almost symmetrical) with the slight tendency to right-branching:

Fiction:	(L/R)=0.974
Scientific texts:	(L/R)=0.984
Poetry:	(L/R)=0.983
Spoken speech:	(L/R)=1.310

Therefore, written Russian in this aspect leans towards the mirror symmetry, while everyday spoken Russian is left asymmetrical [23].

Amazingly, the difference in branching preference between written and spoken Russian is even greater than that between different languages. Thus, we may claim that in this aspect the difference between written and spoken Russian is larger than that between written Russian and written English.

The other syntactic studies made on the ORD data include the analysis of repetitions, interruptions, self-corrections, "plug-in" constructions, and the ways of reporting someone else's speech [24]. Elements of meta-communication that are common for spontaneous speech and that depend on the type of communication and speaker's characteristics have been studied as well.

4 Some Directions for Further Research

Recently, we have started a large sociolinguistic project with an aim to analyze special characteristics of everyday Russian used by different social groups, and to reveal how the language actually functions and what modifications does it have in a nowadays society. Speech of the major social groups of a contemporary Russian city (age-, gender-, professional-related, etc.) has to be analyzed on different linguistic levels in regard to social information about the speakers. In light of this task, it has became necessary to extend the volume of speech data gathered from particular social groupings [25] and to make correspondent adaptation of the corpus itself. Thus, "one day of speech" recordings are continued in 2015.

The study of speech of different social groups from the population of the second biggest Russian city – St. Petersburg – is to be conducted on phonetic, lexical, morphological and syntactic levels. For example, the following parameters are to be analyzed on phonetic level: (a) temporal characteristics

of speech (overall speech tempo, duration of speech elements, typical rhythmic structures); (b) phonetic realization of frequently reduced forms, discursive markers and fillers; and (c) prosodic models for particular types of utterances. Regular studies of intonation have not been earlier conducted on the ORD data.

As for syntactic studies, it is planned to carry out syntactic analysis for the following parameters: (a) linear word order of verbal and noun syntactic groups, (b) syntactic complexity of spoken utterances (e.g., height and width of linearized trees, left-branching structures vs. right-branching ones, syntactic discontinuity) [26], (c) specific syntactic phenomena of spontaneous speech (parcellation, ellipsis, breaks, incompleteness, self-correction, etc.), and (d) the usage of syntactic markers (prepositions, conjunctions, introductory words, etc.).

Besides, it is planned to conduct studies of paralinguistic phenomena and psycholinguistic studies (dependency of speech characteristics from speaker's psychological type) on all linguistic levels.

In this review, we intentionally skipped the description of lexical and morphological studies made on the ORD data. These investigations are actively performed as well (for example, cf. [11,16,24]) and deserve a separate review.

Sociolinguistic extension of the corpus allows to increase the volume of speech transcripts up to 1 million words during the next 1.5 years. Therefore, the ORD corpus will be a representative resource of everyday spoken Russian, suitable for solving both theoretical and applied linguistics problems.

Acknowledgements. The research is made within the framework of the project "Everyday Russian Language in Different Social Groups" supported by the Russian Scientific Foundation, project # 14-18-02070.

References

1. Grishina, E.: Ustnaja rech v Nacionalnom korpuse russkogo jazyka. Nacionalnyj korpus russkogo jazyka: 2003–2005, pp. 94–110. Indrik Publication, Moscow (2005) (in Russian)
2. Kibrik, A., Podlesskaya, V. (eds.): Rasskazy o snovidenijakh. Korpusnoe issledovanie ustnogo russkogo diskursa. Yazyki slavyanskikh kul'tur, Moscow (2009) (in Russian)
3. Krivnova, O.: Russkij rechevoj korpus RuSpeech. In: Proceedings of the VII International Scientific Conference "Fonetika segodnia", pp. 54–56 (2013)
4. Skrelin, P., Volskaya, N., Kocharov, D., Evgrafova, K. et al.: A fully annotated corpus of Russian speech. In: Proceedings of LREC 2010, pp. 109–112, Malta (2010)
5. Kotov, A., Gopkalo, O.: Russkojazychnyj emocional'nyj korpus: kommunikativnoe vzaimodejstvie v real'nykh emocional'nykh situaciajkh. In: Proceedings of the International Conference "Corpus linguistics-2013", pp. 211–216. St. Petersburg State University, St. Petersburg (2013) (in Russian)
6. Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., Sherstinova, T.: The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 250–257. Springer, Heidelberg (2009)

7. Reference Guide for the British National Corpus. http://www.natcorp.ox.ac.uk/docs/URG.xml
8. Campbell, N.: Speech and expression; the value of a longitudinal corpus. In: Proceedings of LREC 2004, pp. 183–186 (2004)
9. ELAN - Linguistic Annotator. Version 4.9.0. http://www.mpi.nl/corpus/html/elan/
10. Praat: Doing Phonetics by computer. http://www.praat.org
11. Sherstinova, T.: Quantitative data processing in the ORD speech corpus of Russian everyday communication. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.) Text and Language: Structures, Functions, Interrelations, pp. 195–206. Praesens Verlag, Wien (2010)
12. Sherstinova, T.: Russian everyday utterances: the top lists and some statistics. In: Thielemann, N., Kosta, P. (eds.) Approaches to Slavic Interaction. Dialogue Studies, vol. 20, pp. 105–116. John Benjamins Publication Company, Amsterdam/Philadelphia (2013)
13. Stepanova, S.: Speech rate as reflection of speakers social characteristics. In: Thielemann, N., Kosta, P. (eds.) Approaches to Slavic Interaction. Dialogue Studies, vol. 20, pp. 117–129. John Benjamins Publishing Company, Amsterdam/Philadelphia (2013)
14. Metlova, V.: Temp rechi v svobodnoj kommunikacii: sociolingvisticheskij aspekt. Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologija 4(28), pp. 58–65 (2014) (in Russian)
15. Bogdanova, N., Palshina, D.: Reducirovannye formy russkoj rechi (opyt leksikograficheskogo opisanija). In: Proceedings of Sc. Conference "Slovo. Slovar'. Slovesnost': Tekst slovaria i kontekst leksikografii", pp. 491–497. RGPU imeni A. Gerzena, St. Petersburg (2010) (in Russian)
16. Stepanova, S., Asinovsky, A., Ryko, A., Sherstinova, T.: Zvukovaja real'nost' slovoizmenitel'nykh affiksov (po dannym Zvukovogo korpusa russkogo jazyka). In: Proceedings of the International Conference "Dialog 2010", pp. 41–46, Bekasovo (2010) (in Russian)
17. Stepanova, S.: Oslyshki i peresprosy kak baza dlia issedovanija vosprijatija rechi. In: Aktual'nye voprosy teoreticheskoj i prikladnoj fonetiki, pp. 383–397, BukiVedi, Moscow (2014) (in Russian)
18. Bogdanova-Beglarjan, N. (ed.): Zvukovoj korpus kak material dlja analiza russkoj rechi. Chast 1. Chtenie. Pereskaz. Opisanie. Philological Faculty of St. Petersburg State University, St. Petersburg (2013) (in Russian)
19. Sherstinova, T.: Ob izokhronnosti strukturnykh jedinic v spontannoj rechi (k postanovke problemy). In: Asinovsky, A.S., Bogdanova N.V. (eds.) Proceedings of XXXVII International Philological Conference, Issue 23, pp. 109–118. St. Petersburg State University, St. Petersburg (2010) (in Russian)
20. Bogdanova-Beglarian, N., Sherstinova, T., and Kisloshchuk, A.: O ritmoobrazujushchej funkcii diskursivnykh jedinic. Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologija 2(22), pp. 7–17 (2013) (in Russian)
21. Bogdanova-Beglarian, N.: Kto ishchet - vsegda li najdet? (o poiskovoj funkcii verbalnykh khezitativov v russkoj spontannoj rechi). In: Proceedings of the International Conference "Dialog-2013", pp. 125–136 (2013) (in Russian)
22. Lapteva, O.A.: Russkij razgovornyj sintaksis. Nauka, Moscow (1976) (in Russian)
23. Martynenko, G.: Sintaksis zhivoj spontannoj rechi: simmetrija linejnykh poriadkov. In: Proceedings of the International Conference "Corpus linguistics-2015" pp. 307–314 (2015) (in Russian)

24. Bogdanova-Beglarjan, N. (ed.): Zvukovoj korpus kak material dlja analiza russkoj rechi. Chast 2. Teoreticheskie i prakticheskie aspekty analiza. Vol. 1. O nekotorykh osobennostjakh ustnoj spontannoj rechi raznogo tipa. Zvukovoj korpus kak material dlja prepodavanija russkogo jazyka v inostrannoj auditorii. Philological Faculty of St. Petersburg State University, St. Petersburg (2014) (in Russian)
25. Baeva, E.M.: O sposobax sociolingvisticheskoj balansirovki ustnogo korpusa (na primere “Odnogo rechevogo dn’a”). Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia, 4(28), pp. 48–57 (2014) (in Russian)
26. Martynenko, G.: Osnovy stilemetrii. Leningrad State University, Leningrad (1988) (in Russian)