Russian Linguistics (2022) 46:77--93
https://doi.org/10.1007/s11185-022-09254-9

Spoken Corpora of Slavic Languages

Nina Dobrushina[1] · **Elena Sokur**[2]

Accepted: 13 June 2022 / Published online: 20 July 2022

Abstract
Spoken corpora are collections of transcribed and annotated audio and /or video recordings of languages or language varieties. The aim of this paper is to present an overview of 51 spoken corpora currently available for Slavic languages and dialects, in particular Belarusian, Bulgarian, Croatian, Czech, Polish, Russian, Slovak, Slovenian, Trasianka, Ukrainian/Rusyn. We identify three groups of corpora according to the type of lect: corpora of standard languages (spoken mainly in an urban environment and existing in both written and oral form), dialects (spoken mainly in a rural environment and unwritten), and bilingual varieties (we call *bilingual* varieties spoken as L2 by people with different L1 languages, as well as all varieties that evolved in a multilingual environment). We survey the corpora in terms of text registers, transcription, and principles of linguistic and extralinguistic annotation. In conclusion, we suggest a list of features that linguists should take into consideration when developing a spoken corpus. Many spoken corpora are currently being created for various Slavic lects, and their developers may use this overview as a source of information on different designs and solutions.

Аннотация
Устные корпуса – это собрания транскрибированных и аннотированных аудио- и/или видеозаписей текстов на разных языках и диалектах. Статья представляет собой обзор 51 устного корпуса славянских идиомов, в частности белорусского, болгарского, русского, словацкого, словенского, трасянки, украинского/русинского, чешского и хорватского языков. Мы выделяем три группы корпусов в зависимости от типа идиома: корпуса литературных вариантов языка (на которых говорят преимущественно в городах и которые используются на письме), корпуса диалектов (на которых говорят преимущественно в сельской местности и не пишут) и корпуса билингвальных вариантов (на которых говорят как на втором люди, первым языком которых является другой язык, или по каким-то другим причинам находящиеся под сильным влиянием других языков). Мы рассматриваем эти корпуса с точки зрения жанров текстов, типов транскрип-

✉ N. Dobrushina
nina.dobrushina@gmail.com

E. Sokur
elena.o.sokur@gmail.com

[1] Linguistic Convergence Laboratory, University of Hamburg, Überseering 35, Postfach # 29, D-22297 Hamburg, Germany

[2] Linguistic Convergence Laboratory, HSE University, 21/4 Staraya Basmannaya str., 106006, Moscow, Russia

ции и принципов лингвистической и экстралингвистической аннотации. В качестве заключения предлагаются своего рода рекомендации лингвистам, делающим устные корпуса. Поскольку в настоящее время создается много новых устных корпусов, этот обзор может быть использован как источник информации о различных дизайнах и технологических решениях.

1 Introduction[1]

Spoken corpora are "principled collections of electronically available, transcribed and annotated audio and/or video recordings of languages or language varieties" (Ruhi et al., 2014, p. 3, with a reference to Andersen, 2010). While written corpora have become a commonplace and their number is constantly growing, the demand for spoken corpora is still much higher than the supply. The main reason is that, as indicated in (Bermel, 2015), the creation of spoken corpora is technically challenging and presents a lot of problems concerning transcription and annotation. Recording speech in its natural environment may require going to the field, communicating with potential speakers, getting their consent for publication, applying special skills in order to make people talk in front of the recorder. Transcribing recorded data, in its turn, is time-consuming and requires special training and a good understanding of the language variety in question.

Meanwhile, corpora of spoken speech are the only feasible option for most language varieties because they are unwritten. This concerns not only minority languages, but also most varieties of large languages, such as vernaculars, dialects, heritage languages and L2 speech, which exist only in the form of face-to-face oral interaction, while writing is regularly used almost exclusively for standard varieties. Therefore, any decent documentation of non-written language implies creating a spoken corpus.

On top of that, spoken corpora are required for various tasks in computational linguistics such as speech recognition. This technology is currently available almost exclusively for the standard varieties of major languages, though developers are seeking to expand the pool of recognizable lects (Arts et al., 2021; Partanen et al., 2020).

These factors make spoken corpora desirable and essential resources. However, despite their potential usability and value, spoken corpora still have not occupied their niche in linguistic analysis. They are rare and far from perfect. If the researcher is lucky enough to find a spoken corpus of the language variety she needs, the recourse to this corpus might return disappointment. Spoken corpora are much smaller in size than written ones, they usually do not offer enough data to generalize about forms or constructions, and they are often very inconvenient for the users.

In this paper, we aim to present an overview of spoken corpora currently available for Slavic languages and dialects, in hope that considering actual practices for the design of spoken corpora might inspire the creation of new ones. We do not consider electronic collections of texts in public repositories, such as CLARIN Virtual Language Observatory https://vlo.clarin.eu/, ELRA http://www.elra.info/en/, or Wiki of the Association for Computational Linguistics https://aclweb.org/aclwiki/List_of_resources_by_language. They are mostly archived, non-searchable collections of texts, and while they are very valuable, they are not adapted for the use by linguists.

At the same time, we were not very strict in selecting corpora for our database. We wanted to cover as many varieties of Slavic languages as possible and make various initiatives in

[1] Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

Slavic linguistics more visible. Some of the corpora we found are very small, and some are not very user-friendly. We looked for the corpora that are publicly available online, presenting spontaneous speech (which is itself a controversial notion, see discussion in Sect. 3.2), preferably searchable and with text-aligned or at least accessible audio files. Many of them have very limited search options. Some others, such as Rureg, the Acoustic Data Base of Russian Regional Speech, are only partly transcribed. Some provide audio only for a part of the transcriptions (Spoken Slovak Corpus). We included several corpora which do not provide audio at all, but contain only spoken data, such as the Tomsk Dialect Corpus. We did not include the corpora where the transcriptions of oral texts constitute only a small portion of the texts, and are not provided with audio, such as the Corpus of Silesian Language.

We discuss the content, annotation and metadata of the corpora. We show that spoken corpora as online time-aligned databases of spoken speech that are accessible for direct search are not numerous, but their number is growing rapidly. Since many spoken corpora are currently being developed for various Slavic lects, this overview will provide researchers with examples of different designs and solutions.

The paper is arranged as follows. In Sect. 2, we present a brief overview of spoken Slavic corpora which we were able to find. Section 3 groups corpora on the ground of their content, namely the kind of texts they represent. Section 4 discusses the issue of representation of spoken speech in the written form and shows different solutions found in Slavic corpora. In Sect. 5, the types of text annotation are listed. Section 6 considers the perspectives of creation of spoken corpora.

2 Spoken Slavic corpora: the overview

We found 51 Slavic corpora of various size and capacity. Our list is not and cannot be comprehensive. First, because there are many local initiatives in this domain, and we keep finding new resources (some were pointed out to us by the reviewers, to whom we are very grateful). Second, because new corpora are constantly emerging, and this makes this overview even more urgent – the exchange of experience should be ongoing.

These corpora cover eight Slavic languages: Russian, Czech, Slovak, Polish, Slovenian, Ukrainian/Rusyn,[2] Bulgarian and Croatian; there are also some Belarusian texts included in TriMCo corpus (together with several other lects of the Baltic-Slavic contact zone) and a corpus of Trasianka, a mixed Belarusian-Russian lect. The appendix[3] contains a list of 51 corpora with their characteristics in terms of size, annotation, availability of audio, idiom type and link to the web resource.

The largest corpus is the Corpus of Spoken Slovak (Rusko & Garabík, 2007); its size equals 6.6 million tokens. The corpus includes both colloquial and public formal speech (the proportions are not mentioned). There is also a very large collection of spoken colloquial speech offered by the corpus of spoken Czech *ORAL Corpus* (5 million tokens).

Five corpora do not provide audio records at all, such as the Spoken Russian National Corpus. Others are at least to some extent multimodal, that is, they contain both audio and a transcript. Some corpora, such as the Corpus of Dialects of the Slovak National Corpus, are not lemmatised or morphologically annotated; users can browse the corpus by searching

[2] Rusyn is a variety that is historically and sociolinguistically closely linked to Ukrainian, but considered a language of its own by many scholars.

[3] Appendix is available at https://github.com/LingConLab/Spoken_slavic_corpora.

for a word or using CQL (Contextual Query Language, a formal language aimed at making queries human readable and writable).

Overall, the most represented language is Russian; there are 30 spoken corpora of various Russian varieties. Russian also has the best coverage in terms of the types of lects, as we will show in the next section.

The spoken corpora surveyed in this paper represent spontaneous and semi-spontaneous speech, which includes interviews with the researcher, staged narratives and dialogues, or speech from spoken media (such as films). In the next section, we will discuss the types of texts available in spoken corpora in more detail.

3 Corpora according to the type of texts

The creation of spoken corpora can be inspired by different motivations which define their content. Some corpora were conceived as national projects and aim at being representative of various types of speech and various text genres. Some others emerged as byproducts of dialectological or ethnographic studies of one region or were designed for a particular investigation and thus represent a very specific lect. Below we will consider spoken corpora according to the type of lect (3.1) and to the text registers (3.2).

3.1 The types of lect

We roughly identify three groups of corpora according to the type of lect: corpora of standard languages (which are spoken mainly in cities and exist in written as well as in oral form), dialects (spoken mainly in villages and not written), and bilingual varieties (this includes varieties spoken as L2 by people with a different language as L1 and all varieties that evolved in a multilingual environment). Some cases do not fit into this classification, such as the corpus of Belarusian-Russian mixed speech (Trasianka), which is unwritten and spoken in cities as well as in rural regions and was influenced by two closely related languages (Hentschel, 2014).

Corpora of all three types are available only for Russian. We found seven corpora of Russian standard speech, fifteen dialect corpora, and eight bilingual corpora. Ukrainian/Rusyn and Croatian have corpora of the standard language only. Slovak, Polish, Czech, Slovenian and Bulgarian have both standard and dialect corpora. For two languages we found spoken corpora of bilingual varieties – Polish and Russian.

Dialect corpora may consist of texts from different dialects, or be devoted to one particular region or even one particular village. Corpora of the first type are the Corpus of Dialects of the Slovak National Corpus (about 25 dialects, including those transitional to Czech, Polish and Rusyn – Gajdošová et al., 2015) and the Dialect corpus of the National Russian corpus (texts from more than twenty different parts of Russia, Letuchij, 2009). The National Russian corpus allows the user to limit the search query to a particular area. Corpora of one dialect can be exemplified by the Spisz Dialect Corpus (a collection of texts documenting the speech of inhabitants of the Polish Spisz region) or the Corpus of the Opochetsky dialect (a northern dialect of Russian). The Spisz Dialect Corpus covers 15 villages and allows filtering texts according to the village where it was recorded.

Dialect corpora are a valuable source of historical and anthropological information. They usually contain life stories which illuminate the history of regions, their culture, and the fate of their residents and families. Some dialect corpora are the products of shared efforts of linguists and anthropologists, such as the Corpus of Spiridonova Buda dialect (Southern

Russian dialect). It consists of interviews on topics related to various aspects of traditional peasant culture, notably mythology, ritualism, folklore and oral history, conducted by a group of anthropologists led by A. B. Moroz in 2017. In 2018, the material was transformed into a searchable spoken corpus by linguists from HSE University.

Another important usage of dialect corpora are quantitative variationist studies, which are difficult to carry out when the texts are not equipped with a search engine. Recent examples of variationist studies based on dialect corpora are the papers by Daniel et al. (2019), based on the Ustja River Basin corpus, and Ter-Avanesova and Daniel (2022) on second genitive in Russian, which uses several dialect corpora of Russian.

The bilingual corpora contain data recorded from speakers of Slavic languages who are strongly influenced by another language. This includes speakers who have a Slavic language as a heritage and/or family language (such as Polish speakers in Germany, see below), bilingual speakers of minority languages who are almost equally proficient in two languages (younger generations of Russian speakers in Daghestanian villages, see below), monolingual speakers of Slavic languages whose family language was non-Slavic (some speakers of Russian in Karelia), and so forth.

For example, the Hamburg Corpus of Polish in Germany contains recordings of bilingual Polish-German speakers currently living in Germany. This corpus was created for a project aimed at describing contact-induced changes in the speech of German Poles. The speakers for the Hamburg corpus were selected on the basis of the aims of the project. As follows from Czachór, 2012, they examined two groups of bilingual speakers according to their personal story of language acquisition and their age at the time of emigration. The first group included participants who acquired Polish as an L1 without instruction, i.e. in natural acquisitional settings within the family, and never attended a Polish school. The second group included speakers who moved to Germany at a later age (>16), after finishing secondary school or even university in Poland. The texts for this project were semi-spontaneous. They consist of interviews addressing the following topics: (1) the participant's best or last holidays; (2) their daily routine and route to work; and (3) their imagination of the world in the year 3000; in addition, participants were asked to describe picture stories (Czachór, 2012).

In contrast to the Hamburg Corpus of Polish in Germany, the available bilingual corpora of Russian were not designed for some specific research project. The largest corpus (376,717 tokens, 102 speakers in April 2022) is a collection of sociolinguistic interviews conducted in Daghestan, a multilingual republic of Russia. All of the texts in this corpus are dialogues with researchers from Moscow. The corpus is a side project of a study of Daghestanian multilingualism. The sample of speakers was thus not planned ahead. At the present time, it comprises the residents of 30 highland villages plus the town city of Makhachkala, whose years of birth are between 1920 and 2000. Bilingual corpora of Russian spoken by L1 speakers of Chuvash, Bashkir, Karelian, Beserman Udmurt and Roma also present collections of dialogues between researchers and speakers about their life and their language repertoires, but they are much smaller.

A special type of bilingual corpora are collections of texts with code-switching. These corpora contain texts of bilingual speakers alternating between two languages within a discourse. In order to annotate such corpora, the researcher needs to have knowledge of both languages. We are aware of two Slavic corpora of code-switching. The corpus of contact-influenced Russian of Northern Siberia and the Russian Far East was not aimed at documenting code-switching. According to the authors, the corpus is a "by-product" of current language documentation projects (Khomchenkova et al., 2019). It instantiates the Russian speech of speakers of Chukchi, Yakut and Yukaghir, which contains plenty of cases of code-switchings. The Yakut-Russian code-switching corpus, by contrast, was targeted at code-

switchings from the outset. It is based on audio data gathered in Yakutsk during a conversation between young Yakut-Russian bilinguals while playing the board game "Monopoly" (Petukhova & Sokur, 2021). Both corpora have rich annotation for a number of attributes.

Bilingual corpora open the way for the quantitative study of the effects of contact between two languages. For example, the corpus of Daghestanian Russian mentioned above was used to test two hypotheses. The first hypothesis was that the speakers tend to over-use left branching because their native languages are left-branching. In particular, constructions with noun phrases with a genitive modifier in Daghestanian Russian tend to have the genitive on the left more often than it happens in monolingual Russian: *mojej babushki plem'annik* 'the nephew of my grandmother' (*plem'annik mojej babushki* in monolingual Russian) (Naccarato et al., 2021). The second hypothesis concerned the omission of prepositions in Daghestanian Russian (Panova & Philippova, 2021): *mog by institut postupit'* 'could have entered the university' (*mog by v institut postupit'* in monolingual Russian), which might be due to the fact that the languages spoken in Daghestan have no prepositions. In both cases, the corpus allowed the application of statistical methods, and some of the results were different from what was suggested based on anecdotal information (Daniel et al., 2010).

Finally, there are two special corpora which deal with speakers with certain disorders. The first one is the Russian corpus of dream stories. The texts are divided into two groups: the control group includes 60 stories by children and adolescents, and the experimental group contains 69 stories by participants with various neurotic disorders (Kibrik & Podlesskaja, 2009). The second is the Croatian discourse corpus of speakers with aphasia, which was designed to make up for the lack of resources necessary to study the speakers with aphasia in Croatian (Kraljević et al., 2017).

3.2 Speech registers

In some corpora, texts are classified into types with a great granularity. For example, the Spoken Russian National Corpus allows the user to choose between dozens of genres, including discussion, interview, retelling, lecture and sermon (Savchuk, 2005, pp. 79–82). In this overview we will limit ourselves to several main types, or *registers*, such as public and private, prepared and spontaneous texts, monologues and dialogues.

The genres covered by a spoken corpus are to a large extent defined by the choice of lect. Dialectal and bilingual varieties are typically found only in the private domain. Dialect corpora most often contain interviews with researchers but have some monological parts as well, if the speaker is prone to monologues. For example, the corpus of Czech dialects (DIALEKT) is described as containing recordings which are mostly informal in nature, even though many of them were obtained within the structured interview research paradigm: "The majority of the transcribed dialect recordings contain a usually unprepared monologue-type speech taking place in a private domestic environment". The topics focus on the traditional rural way of life, covering agriculture, arts and crafts, local customs and traditions, contemporary events, etc. (Goláňová & Waclawičová, 2019).

Fully spontaneous speech data cannot be made available to the public for ethical reasons. Data from interviews carried out by researchers is probably the best we can get as informal speech, at least if the corpus is to be publicly accessible. An example of truly spontaneous speech recordings is the corpus of Russian "One Speaker's Day", collected and transcribed by researchers from Saint Petersburg. The corpus is not available online exactly due to the sensitivity of the data it contains (Asinovsky et al., 2009). Another such example is the Nijmegen Corpus of Casual Czech, which is also not publicly accessible (Kočková-Amortová et al., 2014).

The corpora of standard, non-dialect speech can be aimed at genre representativeness, which means that they cover different settings, diverse situations of speech, and different degrees of formality, such as the Slovenian corpus GOS (Verdonik et al., 2013). The two biggest spoken corpora of Russian, the Spoken Russian National Corpus and the Multimodal Russian National Corpus, cover both public and private situations of speech (Grishina & Savchuk, 2009; Grishina, 2009). In addition to other genres, the Multimodal Russian National Corpus includes video and audio fragments of films from the 1930s to the 2000s, aligned with transcription. The user can search not only by the spoken text, but also by gestures (e.g. nodding one's head, patting on the shoulder), and the type of speech action (agreement, irony, etc.).

A special semi-spontaneous genre is found in the corpora which use "staged" narratives or dialogues. As one example, a small multichannel corpus "Russian Pear Chats and Stories" consists of conversations between people after watching "The Pear Film", a stimulus movie created by a research group led by Wallace Chafe in the 1970s. The idea of Pear stories is to test how much a simple story will vary from language to language. The very fact that Pear stories are collected in many different languages and dialects contributes to the efficiency of this project.

4 Standard orthography or phonetic transcription?

Transcription is the most difficult issue for building spoken corpora. Oral speech has to be represented in writing in order to make it analyzable and searchable. Transferring spoken language into written form requires solutions which directly affect the kind of research issues that can be addressed using the corpus data.

This problem is especially difficult with regard to dialect and bilingual corpora, because they abound with deviations from the standard (written) language, such as non-standard pronunciation or morphemes. This poses the problem of how to deal with such cases. There are three main strategies found in Slavic spoken corpora: phonetic transcription, standard orthography, or some combination of standard orthography with transcription.

The Multimedia Corpus of Spoken Bulgarian (which is a part of the Spoken Bulgarian Corpus) uses modified orthography instead of standard orthography to reflect some of the features of spontaneous speech. Gestures, mimics, pauses and laughter are also represented in the transcription. Another part of the Spoken Bulgarian Corpus, called Parallel Corpus, presents two types of transcriptions in a parallel format, a two-column view with the normalized transcription to the left and the original transcription which reflects some phonetic and morphological features of the spoken text to the right. In Fig. 1, taken from Tisheva et al. (2018), the red colour highlights the discrepancies between two transcripts (p. 25).

One of the disadvantages of using phonetic transcription or keeping dialect forms is that this leads to partial or even complete absence of morphological and syntactic annotation and thus the unavailability of automated search, as in the Bulgarian corpora mentioned above. Automatic annotation of non-standard speech is a difficult technical issue. In most cases, only automatic annotation tools designed for the standard language are available. We should note, however, that there are ongoing attempts to create systems of automated morphosyntactic tagging for low-resourced languages. In the domain of Slavic linguistics, Scherrer and Rabus (2019) discuss their attempt to apply neural tagging trained on data from related languages to Rusyn, i.e. without using any annotated data from Rusyn itself. The results are quite impressive.

Some dialect corpora include two layers aligned with each other and with the audio data: one layer containing standard orthography, and the second layer containing some other type of transcription.

(В) Естествено , нали за това съм я донесла .	(В) Естествено , нали за това съм я донесла .
(Л . Д-ва) . пие ми са бира / обаче съм на антибиотици и не мога . трябва да изчакам още няколко дена и тогава . абе какъв е този червения код бе дес / къде да го търся / аз гледах / гледах и нищо не видях ...Ф... (гледа етикета на бутилката Кока Кола) . кажи ми кое е / че аз купих и тъй и не можах да разбера къде да се обадя .	(Л . Д-ва) / / пий ми съ биръ / убаче съм на ънтибиотици и ни могъ / / тр'абвъ дъ исчакъм ошти н'акулку денъ и тугас / / абе къкъф е тос червенийъ кот бе дес / къде дъ гу търс'ъ / ас гледъх / гледъх и ништу ни видиах ...Ф... (гледа етикета на бутилката с Кока Кола) / / къжи ми куйе и / чи ас купих и тъй и ни мужах дъ ръзберъ къде дъ съ убад'ъ / /
(В) Не им се връзвай на тези , два милиона деветстотин и колко хиляди от така наречените награди са мелодии и лого за джиесеми .	(В) Не им се връзвай на тези , два милиона деветстотин и колко хиляди от така наречените награди са мелодии и лого за джи ес еми .

Fig. 1 Parallel transcripts in the Spoken Bulgarian Corpus (normalized transcription to the left, original to the right) (Tisheva et al., 2018, p. 25)

Table 2 Transcribing different realizations of the modal adverb *lahko* 'can'

Pronunciation-based transcription	*lahko*	*lohk*	*lah*	*lohk*	*lehko*	*lahku*	*lejko*	*uohk*	*lehku*
Phonetic realization (not transcribed)	lax"ko:	"lO:xk	"la:x	"lO:xk	"lE:xko	lax"ku:	"lE:jkO	"wO:xk	"lE:xku
Standardized transcription	lahko	lahko	lahko	lahko	lahko	lahko	lahko	lahko	lahko

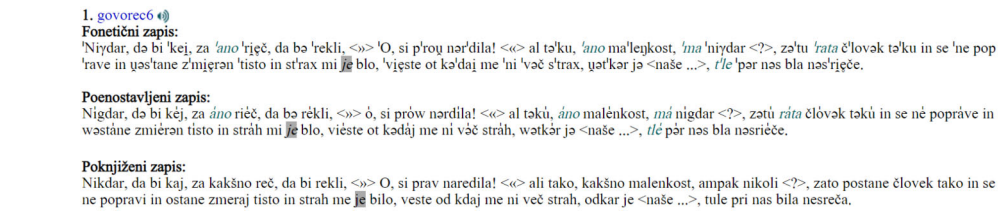
Fig. 2 Pronunciation-based transcription

This method is implemented in the Czech ORTOFON and DIALEKT corpora (Komrsková et al., 2017). The ORTOFON corpus is a spoken corpus of spontaneous everyday communication. Its annotation scheme contains layers of orthographic and phonetic transcriptions. The DIALEKT corpus is similar to ORTOFON, but has a layer of dialectological transcription instead of phonetic transcription. For example, it includes several special symbols for dialect vowels to capture actual pronunciation.

Another example of a mixed transcription system is the Slovenian corpus GOS, which includes standardized and pronunciation-based transcriptions (Verdonik et al., 2013). The authors explain the importance of having a standardized layer of transcription by the need to make it easy to learn for transcribers, and to enable the usage of automatic lemmatization and grammatical annotation. Pronunciation-based transcription is not the same as phonetic transcription; it combines orthography with special symbols for reduced vowels, semi-vowels and some dialect-specific diphthongs. Figure 2 gives an example of how the transcriptions differ (Verdonik et al., 2013, p. 13).

A similar strategy is applied in the corpus of one Slovenian village, Kopriva (Šumenjak, 2013), where a three-fold transcription is applied: a phonetic notation which takes into account all the phonetic characteristics, a simplified record where basic phonetic characteristics of the local speech are kept, and the standard writing system (Fig. 3).

Such a strategy of combining two types of transcriptions has important advantages: the presence of standard orthography enables automatic lemmatization and annotation of grammatical information, while written phonetic data provides the opportunity to search for particular phonetic variants. This is, however, even more time-consuming than phonetic transcription, since it requires creating two levels of transcription instead of one.


1. govorec6
Fonetični zapis:
ˈNiʏdar, də bi ˈkɛj, za *ˈano* ˈrjɛ̨č, da bə ˈrekli, <»> ˈO, si pˈroɥ nəɾˈdila! <«> al təˈku, *ˈano* maˈleŋkost, *ˈma* ˈniʏdar <?>, zəˈtu *ˈrata* čˈlovək təˈku in se ˈne pop ˈrave in ɥəsˈtane zˈmjɛ̨rən ˈtisto in stˈrax mi je blo, ˈvjɛ̨ste ot kəˈdaj me ˈni ˈvəč sˈtrax, ɥətˈkər jə <naše ...>, *tˈle* ˈpər nəs bla nəsˈrjɛ̨če.

Poenostavljeni zapis:
Níɣdar, də bi kéj, za *áno* riéč, da bə rékli, <»> ò, si prów nərdíla! <«> al təkú, *má* níɣdar <?>, zətú *ráta* človək təkú in se nė́ poráve in wəstáne zmiérən tísto in stráh mi je blo, viéste ot kədáj me ni véč stráh, wətkər jə <naše ...>, *tlé* pər nəs bla nəsriéče.

Poknjiženi zapis:
Nikdar, da bi kaj, za kakšno reč, da bi rekli, <»> O, si prav naredila! <«> ali tako, kakšno malenkost, ampak nikoli <?>, zato postane človek tako in se ne popravi in ostane zmeraj tisto in strah me je bilo, veste od kdaj me ni več strah, odkar je <naše ...>, tule pri nas bila nesreča.

Fig. 3 Three levels of transcription in the corpus of Kopriva (a village in Slovenia) (https://jt.upr.si/GOKO/frames-cqp_sl.html)

Most Slavic spoken corpora choose standard orthography. Von Waldenfels et al. (2014) give several reasons in favor of this approach.

First, transcription into the standard language can be done much faster than phonetic transcription. A lot of excellent and extremely valuable data await processing for decades because phonetic transcription is too demanding and time-consuming. Standardization allows to skip the stage of discussions and to make tough decisions about subtle phonetic distinctions which would never satisfy the whole linguistic community.

Second, standard orthography is much less costly in terms of the qualification of the transcribers. To perform phonetic transcription, the transcriber has to be a professional linguist or even dialectologist with an expertise in phonetics and / or dialectology. Standard transcription can be delegated to assistants supervised by senior researchers.

Third, standard orthography makes it possible to use computational tools developed for the standard language, e.g. a morphological tagger for lemmatization and grammatical annotation or a syntactic annotator for analyzing syntactic dependencies in a sentence.

Fourth, the usage of standard orthography facilitates the search process by allowing users to abstract themselves from the variation within and between varieties of the same language.

Fifth, standard orthography is readable for non-linguist users and extends the target audience far beyond linguistics, which is crucial for publicly available resources. This is especially important for dialect corpora, since they can serve as a source of information for anthropologists, historians and the speakers themselves. For example, the authors of the Spisz Dialect Corpus motivate their choice of standard orthography by the intention to make the corpus readable for users who are not familiar with phonetic conventions.

The transcription in standard orthography can differ in the targets of standardization. DIALECT, the corpus of Czech dialects, disregards phone-level differences in word roots in favor of standardized ones, but keeps morphological variation, such as endings of all types of declension (*synoj* vs. standard *synovi* ‘son’ (dative)) and conjugation (*nosijó* vs. standard *nosí* (pl.) ‘they wear’) (Goláňová & Waclawičová, 2019, p. 339). The series of spoken corpora of various Russian dialects (Ustja River Basin, Rogovatka, Spriridonova Buda, Malinino, Opochetsky, Khislavichi, Nekhochi, Lukh and Teza, Upper Pinega and Vyya, Zvenigorod) do not keep morphological deviations, in order to make morphological search possible.

The main disadvantage of transcribing texts in standard orthography is, somewhat paradoxically, the loss of their fundamental property – being non-standard. Standard orthography is justified only under the condition that the corpus gives access to the sound. Aligning the transcript with the original audio on the sentence level makes the spoken corpus suitable for research even if there is no phonetic transcription at all. If audio is available in a user-friendly format, then transcription is only a link between the user and the sound. It allows making queries for morphological categories via grammatical tags and regular expressions in the same way as the standard corpus does, but requires listening to each example. Phonetic

transcription becomes the duty of the user. Such a strategy provides linguists with more corpora than there ever have been (for example, eighteen spoken corpora of different varieties of Russian were launched in the last three years – http://lingconlab.ru/), but it delegates a significant part of the job to the user.

5 Annotation

Corpora usually provide the user with the possibility to search according to certain parameters. In this case texts have to be annotated.

Typical annotation concerns morphology (sometimes including part of speech and lemmatization) and extralinguistic information (metadata).

5.1 Linguistic annotation

The granularity of annotation varies from corpus to corpus. It can include token segmentation, lemmatization, morphological (grammatical), syntactic and discursive analysis, or it can be annotated only with word segmentation (Czech corpora OVM (Otázky Václava Moravce) and Prague DaTabase of Spoken Czech). Besides that, annotation can include special markups of phenomena specific to spoken speech, such as pauses, noises, laughter, gestures, etc.

Most of the corpora provide morphological information, which allows search on various levels:

- words – allows the user to search for a specific word form;
- lemmas – allows the user to find all forms of a word;
- grammatical tags – each word form in the corpus is assigned to grammatical tags which define the values of the grammatical categories the word form has;
- part of speech – if the corpus is annotated for grammatical tags, it may provide information about the part of speech of the tokens.

It is common for search engines to use regular expressions. Standard conventions for regular expressions are, e.g., "*" for any number of syllables, or "+" for any positive number of syllables. Regular expressions happen to be useful when one needs to find tokens following a particular pattern. A regular expression "a.*" will find all words starting with "a" (including "a" if it is a word). One type of interface allowing the use of regular expressions involves SQL (Standard Query Language), such as the Ustja River Basin Corpus.

The Ustja River Basin Corpus is annotated for words, lemmas, morphological categories and parts of speech. Morphological annotation is available to the user in the form of a special interface (Fig. 4).

Syntactic annotation is very rare. Only two of the corpora we reviewed (Prague Dependency Treebank of Spoken Language (PDTSL) 0.5 and The Spoken Slovenian UD Treebank (SST)) contain information about syntax dependencies in spoken speech.

Five Russian corpora (created essentially by the same team) and two Bulgarian corpora provide discursive annotation aimed at studying discourse and prosody. The Multimedia Corpus of Spoken Bulgarian has non-verbal elements (pauses, noise, laughter, etc.), as well as information about the speakers' mimics and gesturing marked in the transcripts. This type of annotation is manual and therefore very time-consuming. Kibrik and Podlesskaja (2003) give the main principles of the annotation of the Russian corpora: they use punctuation marks to indicate important prosodic elements (e.g. a comma or suspension points for a pause, a

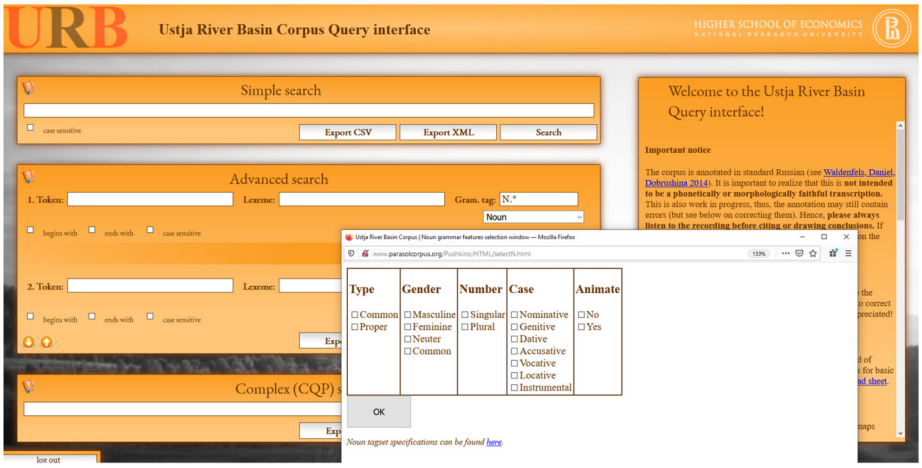


Fig. 4 Interface for grammatical search in the Ustja River Basin Corpus

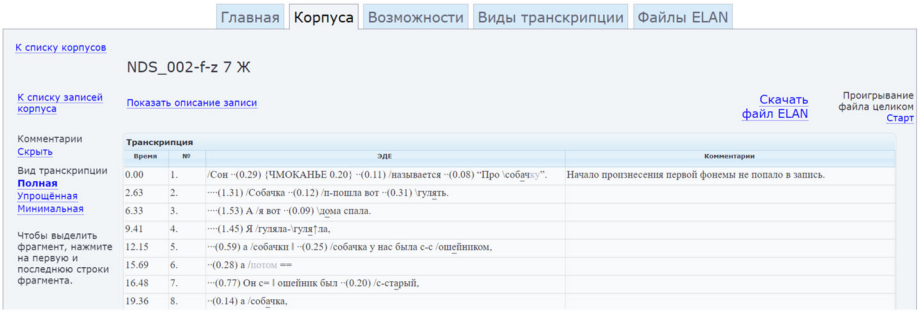


Fig. 5 Discursive annotation in the Dream Stories corpus (Рассказы о сновидениях)

period for a stop), count the duration of pauses, mark the emphasis and its tone, the sound extension, some expressions of emotions (such as laughter or a smile), and sighs. The most important stage of discursive structure analysis is the notation of elementary discursive units (i.e. one clause or a prosodic unit which has a single intonational contour) (Fig. 5).

5.2 Extralinguistic annotation

Extralinguistic information, usually referred to as metadata, is crucial for spoken corpora, since most research tasks which are performed with their help contain a sociolinguistic dimension. Variationist studies often take into account various parameters of the speakers, such as age, gender, place of birth, and education. The Spokes corpus (Conversational Corpus of the Polish language, Pęzik, 2015) provides metadata parameters that allow the user to filter texts based on the age, gender, and education of the speaker. In some special cases, other types of metadata may be included. The Corpus of Russian spoken in Daghestan contains information on the L1 of each speaker, because the area is notoriously multilingual and the peculiarities of Russian speech can depend on the properties of a particular L1. Metadata is sometimes available as additional information, or it can be a part of the parameters available

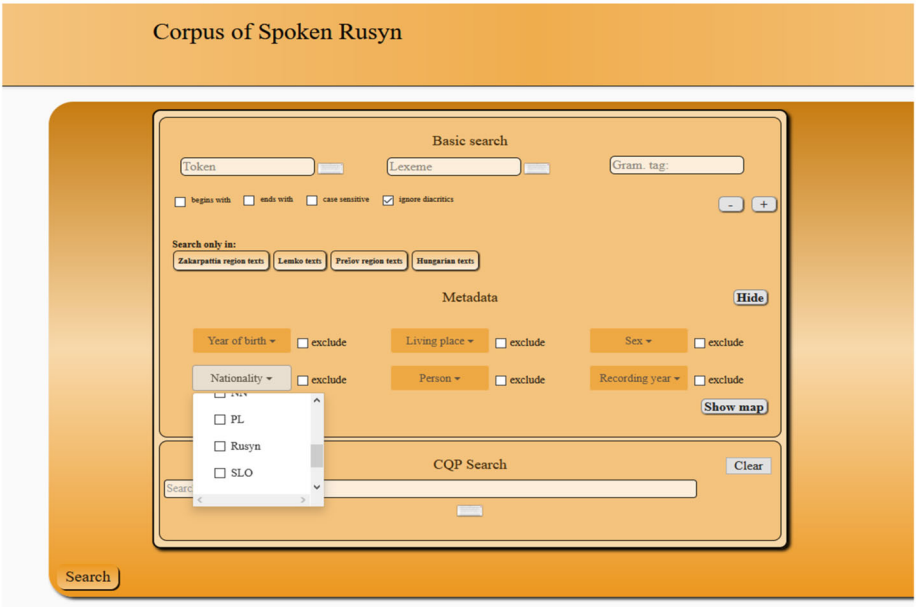


Fig. 6 Interface for metadata in the Corpus of Spoken Rusyn

for search queries, which is more convenient for the user. The former approach is found in The Corpus of Spoken Bulgarian, while the latter is used in the Corpus of Spoken Rusyn (Fig. 6). Apart from information on speakers, the Slovenian GOS corpus provides metadata on the transcriber, the filename of the associated audio, the version and date of the transcription and information about the communicative situation, including discourse type (public, non-public, private), communication channel, type of event, and the place and time of the event.

In a sense, including metadata is even more important than annotating linguistic features. Sometimes dialect texts are recorded without documenting the year of birth and other biographical data of the speaker. Spoken data is valuable in itself, but the absence of metadata reduces their potential usage for research irreversibly. If not dealt with properly from the very beginning, metadata can hardly be retrieved at a later stage.

6 Outlooks

The corpora of Slavic languages reviewed above vary considerably in terms of their search functions and design, from very advanced to basic. Although the design of a corpus depends on the personal preferences of the authors, their aims and their data, the findings of this review allow us to suggest a list of features which linguists would most likely need in order to effectively use a spoken corpus. The reviewed cases of good practice show that a spoken corpus should enable the user to:

- Easily move from external annotation to sound fragments (this might be achieved by chunking the recording and providing a possibility to access a specific chunk via a link);
- Have sound-to-transcript alignment at utterance level;

- Have metadata sensitive queries (the age and gender of the speaker and similar features);
- Have the possibility to download a selection of contexts with links (or some other type of connection) to the sound as csv or a compatible file;
- Have flexible search functions (providing both regex style and layman-friendly options);
- Support regular expressions and CQL (Contextual Query Language, a formal language aimed at making queries human readable and writable), even if a corpus is fully tagged;
- Support multiple word queries (useful for studies of collocations);
- Be simple: as expressed by von Waldenfels and Woźniak (2017), simplicity is the "key issue in spreading corpus use in and beyond the research community".

What was not discussed in this review, is how to store the data so that it survives during a longer time span. To some extent, this is discussed by von Waldenfels and Woźniak (2017) in their paper introducing SpoCo, a system for the web-based search engine of spoken corpora encoded in ELAN. SpoCo is implemented in many Slavic corpora, including the Rusyn Corpus, the Spisz Corpus, and many corpora of varieties of Russian. von Waldenfels and Woźniak (2017) argue that standard formats should be used whenever possible, because today's tools will soon be superseded by more advanced ones, and standard formats will make the migration of the data to new systems less problematic. Storing spoken language data in simple and accessible formats also warrants that valuable recordings and their metadata will be sustained even if the accompanying search engine would cease to function.

Although Slavicists put a lot of effort into the creation of spoken corpora, the field is still far from having its own gold standard. The corpora with audio annotated with a transcription and provided with user-friendly search facilities are not many. There are many Slavic lects that are heavily underrepresented in the domain of spoken corpora, such as Ukrainian, Belarusian, Macedonian and Sorbian.

The potential of spoken corpora is considerable, and, what is especially important, they are not only relevant for linguistics, but also for cultural documentation. The collected narratives and stories from endangered language communities, whatever format they may be in, are of great value to the language community. For endangered dialects, there simply is no agreed-upon writing system which could be a natural vehicle for conveying this kind of material in writing, hence one has no choice but to rely upon audio or audio-visual recordings as the appropriate medium. Spoken corpora are therefore an important opportunity to document these varieties.

Acknowledgements We thank Michael Daniel, Mikhail Kopotev, Olga Lyashevskaya, Samira Verhees, Ruprecht von Waldenfels, and two anonymous reviewers for comments and references.

Funding Note Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest statement On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Andersen, G. (2010). How to use corpus linguistics in sociolinguistics. In A. O'Keefe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 547–562). London/New York: Routledge.

Arts, F., Başkent, D., & Tamati, T. N. (2021). Development and structure of the VariaNTS corpus: a spoken Dutch corpus containing talker and linguistic variability. *Speech Communication*, *127*, 64–72. https://doi.org/10.1016/j.specom.2020.12.006.

Bermel, N. (2015). Corpora and quantitative data in Slavic languages. *Russian Linguistics*, *39*, 275–282. https://proxylibrary.hse.ru:2120/10.1007/s11185-015-9154-5.

Czachór, A. (2012). Corpus of Polish Spoken in Germany. Collecting and analysing written & spoken data for investigating contact-induced change. In T. Schmidt & K. Wörner (Eds.), *Studies on multilingualism: Vol. 14. Multilingual corpora and multilingual corpus analysis* (pp. 153–161). Hamburg: Benjamins.

Daniel, M. et al. (2019). Dialect loss in the Russian North: modeling change across variables. *Language Variation and Change*, *31*(3), 353–376. https://doi.org/10.1017/S0954394519000243.

Daniel, M., Knyazev, S., & Dobrushina, N. (2010). Highlander's Russian: case study in bilingualism and language interference in Central Daghestan. In A. Mustajoki, E. Protassova, & N. Vakhtin (Eds.), *Slavica Helsingiensia: Vol. 40. Russian language in the multilingual world* (pp. 65–93). Helsinki: University of Helsinki.

Goláňová, H., & Waclawičová, M. (2019). The DIALEKT corpus and its possibilities. *Journal of Linguistics/Jazykovedný casopis*, *70*(2), 336–344. https://doi.org/10.2478/jazcas-2019-0063.

Grishina, E. A. (2009). Mul'timedijnyj russkij korpus (MURKO): problemy annotatsii. In V. A. Plungyan (Ed.), *Natsional'nyj korpus russkogo jazyka. Novyje resul'taty i perspektivy* (pp. 175-–214). Saint-Petersburg: Nestor-Istorija.

Grishina, E. A., & Savchuk, S. O. (2009). Korpus ustnyh tekstov v NKRYa: sostav i struktura. In V. A. Plungyan (Ed.), *Natsional'nyj korpus russkogo jazyka. Novyje resul'taty i perspektivy* (pp. 129–149). Saint-Petersburg: Nestor-Istorija.

Hentschel, G. (2014). Belarusian and Russian in the mixed speech of Belarus. In J. Besters-Dilger et al. (Eds.), *Congruence in contact-induced language change* (pp. 93–121). Berlin/Boston: de Gruyter.

Khomchenkova, I. A., Pleshak, P. S., & Stoynova, N. M. (2019). The corpus of contact-influenced Russian of Northern Siberia and the Russian Far East. In V. P. Selegey (Ed.), *Computational linguistics and intellectual technologies: papers from the annual international conference "Dialogue 2019"* (Vol. 18, pp. 276–287). Moscow: RGGU.

Kibrik, A. A., & Podlesskaja, V. I. (2003). K sozdaniju korpusov ustnoj russkoj rechi: printsipy transkribirovanija. *Nauchno-tehnicheskaja informatsija*, *2*(6), 5–11.

Kibrik, A. A., & Podlesskaja, V. I. (2009). Rasskazy o snovidenijah: korpusnoje issledovanije ustnogo russkogo diskursa. In A. A. Kibrik & V. I. Podlesskaja (Eds.), *Night Dream Stories: a corpus study of spoken Russian discourse*. Moscow: Jazyki slavjanskih kul'tur.

Kočková-Amortová, L., Pollák, P., Rajnoha, J., & Ernestus, M. (2014). The Nijmegen corpus of casual Czech. In N. Calzolari (Ed.), *Proceedings of LREC 2014: 9th international conference on language resources and evaluation* (pp. 365–370). Reykjavik, Iceland: European Language Resources Association (ELRA).

Komrsková, Z., Kopřivová, M., Lukeš, D., Poukarová, P., & Goláňová, H. (2017). New spoken corpora of Czech: ORTOFON and DIALEKT. *Journal of Linguistics/Jazykovedný casopis*, *68*(2), 219–228. https://doi.org/10.1515/jazcas-2017-0031.

Kraljević, J. K., Hržica, G., & Lice, K. (2017). CroDA: a Croatian discourse corpus of speakers with aphasia. *Hrvatska revija za rehabilitacijska istraživanja*, *53*(2), 61–71. https://doi.org/10.31299/hrri.53.2.5.

Letuchij, A. B. (2009). Dialektnyj korpus: sostav i osobennosti razmetki. In V. A. Plungyan (Ed.), *Natsional'nyj korpus russkogo jazyka. Novyje resul'taty i perspektivy* (pp. 114–128). SPb.: Nestor-Istorija.

Naccarato, C., Panova, A., & Stoynova, N. (2021). Word-order variation in a contact setting: a corpus-based investigation of Russian spoken in Daghestan. *Language Variation and Change*, *33*(3), 387–411. https://doi.org/10.1017/S095439452100017X.

Panova, A., & Philippova, T. (2021). When a cross-linguistic tendency marries incomplete acquisition: preposition drop in Russian spoken in Daghestan. *International Journal of Bilingualism*, *25*(3), 640–667. https://doi.org/10.1177/1367006921990442.

Partanen, N., Hämäläinen, M., & Klooster, T. (2020). Speech recognition for endangered and extinct Samoyedic languages. In M. L. Nguyen, M. Ch. Luong, & S. Song (Eds.), *Proceedings of the 34th Pacific Asia conference on language, information and computation* (pp. 523–533). Hanoi, Vietnam: Association for Computational Linguistics.

Petukhova, A., & Sokur, E. (2021). Creating a spoken corpus of Yakut-Russian code-switching. In V. P. Selegey (Ed.), *Computational linguistics and intellectual technologies* (pp. 1161–1169). Moscow: RGGU. Supplementary volume.

Ruhi, Ş., Haugh, M., Schmidt, T., & Wörner, K. (2014). Introduction: putting practices in spoken corpora into focus. In Ş. Ruhi, M. Haugh, T. Schmidt, & K. Wörner (Eds.), *Best practices for spoken corpora in linguistic research* (pp. 3–17). Newcastle upon Tyne: Cambridge Scholars Publishing.

Rusko, M., & Garabík, R. (2007). Corpus of spoken Slovak language. In J. Levická & R. Garabík (Eds.), *Computer treatment of Slavic and East European languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2007* (pp. 222–236). Brno: Tribun.

Savchuk, S. O. (2005). Metatekstovaja razmetka v Natsional'nom korpuse russkogo jazyka: bazovyje printsipy i osnovnyje funktsii. In *Natsional'nyj korpus russkogo jazyka: 2003-2005. Rezul'taty i perspektivy* (pp. 62–88). Moscow: Izdatel'stvo Indrik.

Scherrer, Y., & Rabus, A. (2019). Neural morphosyntactic tagging for Rusyn. *Natural Language Engineering*, *25*(5), 633–650. https://doi.org/10.1017/S1351324919000287.

Šumenjak, K. (2013). Priprava gradiva in standardizacija nivojev zapisa za potrebe dialektološkega korpusa GOKO. In A. Žele (Ed.), *Družbena funkcijskost jezika (vidiki, merila, opredelitve)*. (Vol. *32*, pp. 443–449). Ljubljana: Znanstvena založba Filozofske fakultete.

Ter-Avanesova, A., & Daniel, M. (2022). *The second genitive in the history of Russian and across its dialects. Linguistic variation*. https://doi.org/10.1075/lv.21004.ter. Online-first articles.

Tisheva, Y., Dzhonova, M., & Hauge, K. R. (2018). The Corpus of Spoken Bulgarian. Papers of BAS. *Humanities and Social Sciences*, *5*(1), 20–28.

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation*, *47*(4), 1031–1048. https://doi.org/10.1007/s10579-013-9216-5.

Wiemer, B., Kozhanov, K. A., & Erker, A. (2019). Korpus slav'anskih i baltijskih govorov TriMCo: struktura, tseli i primery primenenija. In V. A. Dybo (Ed.), *Balto-slav'ankije issl'edovanija–XX* (pp. 122–143). Moscow: RGGU.

Von Waldenfels, R., & Woźniak, M. (2017). SpoCo – a simple and adaptable web interface for dialect corpora. *Journal for Language Technology and Computational Linguistics*, *31*(1), 155–170.

Von Waldenfels, R., Daniel, M., & Dobrushina, N. (2014). Why standard orthography? Building the Ustya River Basin corpus, an online corpus of a Russian dialect. In V. P. Selegey (Ed.), *Kompjuternaja lingvistika i intelektual'nyje tehnologii* (pp. 720–728). Moscow.

Corpora mentioned in the text

Arkhangelskiy, T. (2020). *Corpus of Russian spoken by the Besermans*. Moscow: Linguistic Convergence Laboratory, HSE University. Available online at URL: http://lingconlab.ru/BesermanRus/, accessed on 19.04.2022.

Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., & Sherstinova, T. (2009). The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation. In V. Matoušek & P. Mautner (Eds.), *International conference on text, speech and dialogue* (pp. 250–257). Berlin: Springer.

Bayda Ivanova, K., Kholodilova, M., Kozhemjakina, A., Romanova, E., Remizova, T., Storozheva, A., Tarasova, N., Zorina, A., Morozova, V., Panova, A., & Dobrushina, N. (2018). *ChuvashRus corpus*. Moscow: Linguistic Convergence Laboratory, NRU HSE. URL: http://lingconlab.ru/ChuvashRus/, accessed on 19.04.2022.

Brehmer, B. (2011). Hamburg Corpus of Polish in Germany (HamCoPoliG). Archived in Hamburger Zentrum für Sprachkorpora. Version 0.2. Publication date 2011-09-02. http://hdl.handle.net/11022/0000-0000-63CE-9.

Corpus of Spoken Slovak. Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. https://korpus.sk/shk_en.html.

Daniel, M., Dobrushina, N., & von Waldenfels, R. (2013–2018a). Govor bassejna Ustji. Korpus severnorusskoj dialektnoj rechi. Bern, Moscow. Electronic resource: www.parasolcorpus.org/Pushkino.

Daniel, M., Dobrushina, N., & von Waldenfels, R. (2013–2018b). The language of the Ustja river basin. A corpus of North Russian dialectal speech. Bern, Moscow. Electronic resource: www.parasolcorpus.org/Pushkino.

Dobrovoljc, K., & Nivre, J. (2016). The universal dependencies treebank of spoken Slovenian. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 1566–1573). Portorož: European Language Resources Association (ELRA).

Dobrushina, N., Daniel, M., Waldenfels, R., Maisak, T., & Panova, A. (2018). *Corpus of Russian spoken in Daghestan*. Moscow: Linguistic Convergence Laboratory, NRU HSE. Available online at http://www.parasolcorpus.org/dagrus/, accessed on 06.04.2022.

Gajdošová, K., Garabík, R., & Šimková, M. (2015). Corpus of dialects of the Slovak National Corpus. Electronic resource: https://korpus.sk/attachments/publications/2015_Gajdo%C5%A1ov%C3%A1_Garab%C3%ADk_%C5%A0imkov%C3%A1_Corpus_of_Dialects_of_the_Slovak_National_Corpus.pdf.

Garder, M., Petrova, N., Moroz, A., Panova, A., & Dobrushina, N. (2018). *Korpus govora sela Spiridonova Buda*. Moscow: Linguistic Convergence Laboratory, NRU HSE. Electronic resource: http://lingconlab.ru/SpiridonovaBuda/.

Goláňová, H., Waclawičová, M., Komrsková, Z., Lukeš, D., Kopřivová, M., & Poukarová, P. (2017). DIALEKT: nářeční korpus, verze 1 z 2. 6. 2017. Ústav Českého národního korpusu FF UK, Praha. Electronic resource: http://www.korpus.cz.

Grochola-Szczepanek, H., Górski, R. L., von Waldenfels, R., & Woźniak, M. (2019). Korpus języka mówionego mieszkańców Spisza. *LingVaria*, *14*(27), 165–180. https://doi.org/10.12797/LV.14.2019.27.11.

Hajič, J., Pajas, P., Ircing, P. et al. (2017). Prague DaTabase of spoken Czech 1.0, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2375.

Hajič, J., Pajas, P., Mareček, D. et al. (2009). Prague Dependency Treebank of spoken Language (PDTSL) 0.5, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0001-4914-D.

Hentschel, G., Zeller, J. P., & Tesch, S. Das Oldenburger Korpus zur weißrussisch-russischen gemischten Rede: OK-WRGR. https://uol.de/ok-wrgr.

Khomchenkova, I., Pleshak, P., & Stoynova, N. The corpus of contact-influenced Russian of Northern Siberia and the Russian Far East. Available online at: http://web-corpora.net/ruscontact/.

Knyazev, S. V. (2021). *Corpus of the Russian dialect spoken in the basins of Upper Pinega and Vyya rivers*. Moscow: Linguistic Convergence Laboratory, HSE University. Available online at URL: http://lingconlab.ru/vaduga/, accessed on 19.04.2022.

Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., & Škarpová, M. (2017a). ORTOFON: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Ústav Českého národního korpusu FF UK, Praha. Electronic resource: http://www.korpus.cz.

Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P., Waclawičová, M., Benešová, L., & Křen, M. (2017b). ORAL: korpus neformální mluvené češtiny. verze 1 z 2.6.2017. Ústav Českého národního korpusu FF UK, Praha. Electronic resource: http://www.korpus.cz.

Kozhanov, K., Khadasevich, A., & Chernykh, A. (2020). *Corpus of Russian spoken by Roma*. Moscow: Linguistic Convergence Laboratory, HSE University. URL: http://lingconlab.ru/romarus/, accessed on 19.04.2022.

Kulik, G. et al. (2018). Korpus Ślōnskij Mōwy (Silesian Language Corpus). https://silling.org/kontext/, accessed 25 April 2022.

Kuvshinskaya, Yu.M. (2020). *Corpus of the Russian dialect spoken in the basins of Lukh and Teza rivers*. Moscow: Linguistic Convergence Laboratory, HSE University. Available at: http://lingconlab.ru/lukhteza/, accessed on 19.04.2022.

Multimedia Corpus of Spoken Bulgarian. http://bgspeech.net/bg/resources/multimediacorpus_en.html.

Panova, A. (2021). *Corpus of Russian spoken in Zvenigorod*. Moscow: Linguistic Convergence Laboratory, HSE University. Available online at URL: http://lingconlab.ru/zvenigorod/, accessed on 19.04.2022

Petukhova, A., & Sokur, E. (2021). *Yakut-Russian corpus of code-switching*. Moscow: Linguistic Convergence Laboratory, NRU HSE. Available online at: http://lingconlab.ru/cs_yakut, accessed on 28.05.2021.

Rasskazy o snovidenijax i drugie korpusa zvučaščej rechi. [Night dream stories and other collections of spoken discourse]. Prosodically annotated corpus of spoken Russian (PrACS-Russ). Pilot version. Online: http://spokencorpora.ru.

Ronko, R., Volf, E., Grebyonkina, M., Ershova, M., Okhapkina, A., Khadasevich, A., & Morozova, V. (2019a). *Corpus of Opochetsky dialects*. Moscow: Linguistic Convergence Laboratory, HSE University; V.V. Vinogradov Russian Language Institute Russian Academy of Science. Available online at URL: https://lingconlab.ru/opochka, accessed on 19.04.2022.

Ronko, R. V., Wolf, E. A., Grebjonkina, M. Yu., Ershova, M. Yu., Okhapkina, A. V., Khadasevich, A. S., & Morozova, V. A. (2019b). *Korpus Opochetskih govorov*. Moscow: Linguistic Convergence Laboratory, NRU HSE; Institut russkogo jazyka im. V. V. Vinogradova RAS. Electronic resource: http://lingconlab.ru/luzhnikovo, accessed on 11.11.2019.

Ryko, A. I., & Spiricheva, M. V. (2020). *Corpus of the Russian dialect spoken in Khislavichi district*. Moscow: Linguistic Convergence Laboratory, HSE University. Available online at URL: http://lingconlab.ru/khislavichi/, accessed on 19.04.2022

Sappok, Ch., et al. RuReg: Russian Regions Acoustic Speech Database. http://rureg.ub.rub.de/.

Šmídl, L., & Pražák, A. (2013). OVM – Otázky Václava Moravce, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-000D-EC98-3.

Sokur, E., & Morozova, V. (2019). *Corpus of Russian spoken in Bashkortostan*. Moscow: Linguistic Convergence Laboratory NRU HSE. URL: https://lingconlab.ru/BashkirRus/, accessed on 19.04.2022.

Pęzik, P. (2015). Spokes – a search and exploration service for conversational corpus data. In *Selected papers from the CLARIN 2014 conference*, October 24–25, 2014, Soesterberg, The Netherlands. (pp. 99–109). Linköping: Linköping University Electronic Press, Linköpings universitet.

Šumenjak, K. Korpus GOKO (Govorni korpus Koprive na Krasu). https://jt.upr.si/GOKO/frames-cqp_sl.html.

Ter-Avanesova, A. V., Balabin, F. A., Dyachenko, S. V., Malysheva, A. V., Panova, A. B., & Morozova, V. A. (2019). *Corpus of the Malinino dialect*. Moscow: Linguistic Convergence Laboratory, NRU HSE; V.V. Vinogradov Russian Language Institute of the Russian Academy of Science. Available online at URL: https://lingconlab.ru/malinino/, accessed on 19.04.2022.

Ter-Avanesova, A. V., Dyachenko, S. V., Kolesnikova, E. V., Malysheva, A. V., Ignatenko, D. I., Panova, A. B., & Dobrushina, N. R. (2018). *Corpus of Rogovatka dialect*. Moscow: Linguistic Convergence Laboratory, NRU HSE. URL: http://lingconlab.ru/rogovatka/, accessed on 19.04.2022.

Ter-Avanesova, A. V., Dyachenko, S. V., Korpechkova, E. V., Malysheva, A. V., Pekunova, I. S., & Tolstaya, M. N. (2020). *Corpus of the Nekhochi dialect*. Moscow: Linguistic Convergence Laboratory HSE University, V.V. Vinogradov Russian Language Institute of the Russian Academy of Science, Institute of Slavic Studies of the Russian Academy of Science. Available online at URL: http://lingconlab.ru/nekhochi/, accessed on 19.04.2022.

The "Russian Pear Chats & Stories" corpus (RUPEX). Pilot version. Online: https://multidiscourse.ru/main/?en=1.

The Russian National Corpus. Dialectal corpus (2003–2021). https://ruscorpora.ru/new/en/search-dialect.html.

The Russian National Corpus. Spoken corpus (2003–2021). https://ruscorpora.ru/new/en/search-spoken.html.

Tomskij dialektnij korpus 2.0. [Electronic resource]. Laboratorija obschej i sibirskoj leksikografii NI TGU. http://losl.tsu.ru/?q=losl_search, accessed on 18.04.2022.

Verdonik, D., Potočnik, T., Sepesy Maučec, M., & Erjavec, T. (2018). Spoken corpus Gos VideoLectures 3.0 (transcription), Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1190.

Vlakhov, A., Aksenova, A., Aleksandrova, P., Dyomina, V., Mamonova, T., Nedaykhleb, P., Oleinik, D., Pinaeva, M., Ryzhkov, A., Sazonova, V., Smirnova, A., Terekhina, L., Timoshina, A., & Scheglova, E. (2020). *Karelian Russian corpus*. Moscow: Linguistic Convergence Laboratory, HSE University. Available online at URL: https://lingconlab.ru/karelrus/, accessed on 19.04.2022.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2013). Spoken corpus Gos 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1040.