Papers of BAS
Humanities and Social Sciences
Vol. 5, 2018, No. 1

The Corpus of Spoken Bulgarian

Yovka Tisheva, Marina Dzhonova, Kjetil Rå Hauge

Abstract. The paper presents the features of a corpus compiled exclusively of data on spoken Bulgarian. The strategy for building the corpus is motivated by the features of spoken communication and the aim is to preserve the characteristics of spoken language when transcribed into a text file. The transcription and annotation systems provide a clear and accessible representation of language varieties used in formal and informal contexts. The pragmatic and socio-cultural information given about the speakers and the settings makes the corpus data applicable also in a wider field of humanitarian studies.

Keywords: Corpus, spoken communication, spoken Bulgarian, transcription

1. Introduction

In the quest for a viable basis for theoretical and applied studies, the corpus, alongside an appropriate theoretical framework and methodology, has become an integral part of linguistic research. We now have at our disposal a variety of language resources for Bulgarian: large (representative) or small (specialized) corpora; monolingual, bi- or polylingual corpora, parallel corpora, corpora from written or spoken texts, automatically constructed corpora from already existing written texts (i.e., from printed publications) and corpora compiled from audio and video recordings.

In this paper we present the features of corpora available online and representing spoken Bulgarian as used in formal and informal communication: The Corpus of Spoken Bulgarian (CSBg, available at http://www.bgspeech.net/). This corpus is being developed at the Faculty of Slavic Studies, Sofia University, by researchers from the Department of Bulgarian language. Access is free and available at bgspeech.net. The design and structure of this resource are linked both to the specifics of spoken language, face-to-face communication, literary standard and non-standard pronunciation, dialect features, etc., as well as to the conventions for recording and transcribing spoken language.

A linguistic corpus is, according to David Crystal "A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language (corpus linguistics)" (Crystal 2008 (1980), 117) and the CSBg strives to follow these criteria in the collecting of data and structuring of the resource. (Some definitions add "usually stored in a computer database" (McCarthy 2004, 4), and this is among the future plans for CSBg, but see below for alternative ways of accessing parts of the corpus.) The definitions applied on work on corpora for Bulgarian also tend to integrate the core features of language corpus as "large collection of language samples, suitable for computer processing and selected according to specific (linguistic) criteria, so that it represents an adequate language model" (Koeva 2010, 9).

The CSBg collaborators compile, maintain and provide access to systematically organized partially annotated texts in a digital format, selected to represent the main features of spoken Bulgarian. The uniqueness of corpora of this type is related to the specifics of the language data (spoken language, communication channel, modality, etc.) and the way they were collected and processed. What distinguishes CSBg from the more well-known corpora of contemporary Bulgarian is the nature of the data it includes. The text files in CSBg are transcripts of audio or video recordings of oral communication. In this sense, the written texts in the corpus are secondary to the original speech acts.

2. Background

Practice around the world shows that oral communication data are collected into separate, self-contained corpora of varying volume and degree of representativeness or included in representative national corpora as a sub-corpus under the main database of written texts. In the British National Corpus (http://www.natcorp.ox.ac.uk), the spoken part constitutes 10%. Interestingly, one of the very first large spoken language corpora, The Bergen Corpus of London Teenage Language (COLT) is now a part of the British National Corpus. Oral texts presenting varieties of English are also available in the Australian National Corpus (https://www.ausnc.org.au), and in the East African, Indian, and New Zealand components of the International Corpus of English (http://ice-corpora.net/ice), etc.

For Slavic languages, the practice of the Czech National Corpus (http://www.korpus.cz) is well known. This corpus includes five separate subgroups presenting oral communication in various communicative situations. The Corpus of Spoken Russian is a subpart of the Russian National Corpus (http://www.ruscorpora.ru) and includes recordings of public and spontaneous speech and transcripts from Russian movies.

The resource under consideration here is not part of a larger corpus. The largest representative online corpus of Bulgarian, Bulgarian National Corpus (BulNC), consists of written texts "The transcripts of spoken data constitute only 1% of the total number of texts in the BulNC and represent two domains - Bulgarian parliament debates and lectures on homeopathy" (Koeva, Blagoeva,

Kolkovska 2010, 3680). The Bulgarian Reference Corpus, a resource developed by BulTreeBank project (http://bultreebank.org), includes public and political speeches, but standard spelling was used to compile the text files, which limits the usefulness of the resource.

There are two main types of definitions of *spoken* corpora. The first group follows Chafe's idea (Chafe 1982) that corpus is to display the basic characteristics of a speech used by the members of a speech community in a most faithful manner, thus the data must come from informal unprepared/unscripted oral communication. Informal spontaneous speech is important language variety and the data obtained will reveal the dynamics of the language system. On the other hand, spoken corpora are defined as collections of any language data whose original presentation is in oral form (for an extended discussion see Sinclair 1996). Following the fieldwork tradition of Bulgarian dialectology and sociolinguistics, in the initial phase of collecting data for spoken Bulgarian, everyday informal dialogues were recorded in order to obtain data for colloquial discourse features, informal registers, and deviations from standard pronunciation and grammar. The spoken data collected at this stage showed the characteristics of only one communicative situation and thus CSBg was not totally representative of oral communication, only for a small segment of it. It was also difficult to acquire and process (transcribe and digitize) the recordings and determine the level of spontaneity of the speakers. In order to increase the representativeness of the corpus, recordings and transcripts of TV talk shows, interviews from national and regional radio stations, and samples of academic communication were included. The notion of *spoken* corpus naturally evolved, and the term is used here in a broader sense to describe a collection of transcribed and annotated audio and video recordings representing language varieties used by speakers in oral communication.

3. Features of CSBg

The specificity of empirical data in a corpus of spoken language determines the essential differences in its compilation and structuring, compared to corpora of written texts. Since a corpus of spoken language is generated from audio and video recordings, the most significant features of spoken (oral) language such as elisions, ellipsis, repetitions, doubling, fragmentation, etc., must be represented in the transcripts.

Compliance with the literary orthographic and grammar norms is mandatory in written communication, but in oral communication norms of literary pronunciation, as well as grammatical and lexical norms, are less strict and their application varies in different situations.

The complex of linguistic means used in oral communication follows the basic features of the national (official) language because it is part of it. Nevertheless, its phonetics and grammar do not fully follow all the specifics of the written literary language, nor do they comply with any existing dialect norm. Some of the peculiarities that are noted in corpora of transcribed oral speech are elisions, ellipses, abbreviated forms or phrases, overlapping utterances, incomplete

utterances, repetition of constituents or phrases, colloquial constructions, pragmatic markers, and discourse markers.

The transcripts also give information about the paralinguistic means used by the speakers (pauses, gestures, mimics, fillers, etc.), as these are an integral part of oral communication. Along with linguistic information, a mandatory condition for a corpus to be representative is to include meta-information (socio-demographic features of the speakers, information about the recording, etc.).

The special features of speech also call for a specific approach to designing this type of corpus. While the first level of annotation in corpora of written language are lemmatization and part-of-speech analysis, in corpora of spoken language partial syntactic and pragmatic annotation is carried out as an integral part of the transcription of the recordings, as this is necessary in order to determine the boundaries between the individual utterances and thus to organize the transcription into a dialogical form. The transcriber notes cases of simultaneous speaking, pauses, and overlapping, as well as non-verbal information, and the communicative status of the utterances, during the initial processing of the texts. This also applies to the metadata that accompany every transcript - information about speakers and recordings is also provided.

The selected system for transcribing spoken language employs the normative orthography, but only to a certain degree. If the aim when compiling corpora of spoken language is to maintain the specifics of that spoken language, then there will always be deviations from the norm. When transcribing texts representing Bulgarian dialects, transcription system that renders the full set of dialectal distinctions is traditionally applyed (e.g. *Bulgarian Dialectology as Living Tradition* available at: bulgariandialectology.org/principles-data-presentation). In the transcripts of data collected for sociolinguistic investigations, the deviations from orthoepy as well as dialectal features are marked (e.g. *Corpus of Krasimira Alexova* available at: http://folk.uio.no/kjetilrh/bulg/Aleksova/Transliteration.html). Type of orthographic transcription is used in CSBg transcripts and standard orthography is not adhered in the cases of non-standard pronunciation (e.g. vowel reduction, yat reflexes, stress alternations, elisions, verbal endings, etc.). The typical features of spoken texts as overlaps, repairs, pauses, etc. are also marked. The transcription system is design to be easy to learn and apply for transcribers and comprehensible for the CSBg users.

Unlike written language, where there are clear boundaries between sentences, marked by punctuation, spoken language consists of utterances and turns that are connected by intonation but usually do not adhere to the structural models of written language. For example, the use of punctuation marks to define the communicative status of the utterance in a corpus of spoken language will be much more limited - in the transcripts only question marks and exclamation marks are used in order to show the communicative status of the utterance.

Speech is essentially a dialogue and therefore the transcripts should also be structured in a corresponding format. Monologues are relatively rare and occur mainly in specific situations or speech genres, i.e., lectures, declarations, sermons, and similar. The texts of the corpus represent dialogues in real communicative situations.

When compiling the corpus, it is important to preserve the original form of spoken communication, with consecutive turns by two or more speakers. The text in the transcripts is organized in turns, utterances, and phrases. A turn consists of at least one utterance uttered by one of the speakers. Spontaneous speech naturally shows examples both of turn-taking and overlapping utterances.

CSBg started as a collection of plain-text transcripts of audio recordings of spontaneous speech in non-formal situations. Later on, an XML standard for annotation of spoken corpora was adopted (https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.htm). With the growing collection of audio and video recordings and the possibility to use special tools for compiling a multimodal corpus (see Pavlidou, Kapellidi, Karafoti 2014 for description of different platforms for spoken data processing), the texts in new transcripts was aligned with the sound track of the corresponding audio or video recording. For this purpose the freely available EXMARaLDA Partitur-Editor (https://exmaralda.org/en/partitur-editor-en) was used. This platform allows the transcriber to represent different types of information: metadata relevant to the recording, the situation and the speakers and to the transcribed speech as well as non-verbal information concerning gestures and mimics, pauses, and overlapping. Transcripts made in EXMARaLDA are exportable to XML format, thus allowing publishing the transcription aligned with the original sound recording.

The first part of the corpus consists of transcriptions in text format, mostly of audio recordings of non-formal communication between friends, associates, and relatives. The recordings from formal communication are from media and school communication with a high degree of spontaneity. These transcripts also include meta-information about the recording and the speakers (their education, age, occupation, etc.). Non-verbal information, such as pauses, gestures, and mimics, is also included in the text transcripts. The metadata contain information concerning the communication channel (radio, television, face-to-face, Internet), the degree of preparedness of the speech (spontaneous, prepared), the domain (informal, formal, mass media, politics, business, academic), and the recording (audio, video; if broadcast: name; date and hour of recording/broadcast). This information allows classifying the transcripts and to study how the situation, the interrelation between the speakers and the speakers' characteristics determine the structure of the dialogue and the occurrence of pragmatic and discourse markers. Table 1 presents a simplified description of the data, showing the type of recording as well as the character count and word count in each part of the corpus. Transcripts are described as formal or informal according to the type of communication.

Table 1. Transcripts in text format

Source	**Domain**	**Words**	**Characters**	**Duration**
audio	informal	46,181	251,215	6:00:00
audio	formal	13,290	79,215	1:51:00
TOTAL		**59,471**	**330,430**	**7:51:00**

(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

In a separate normalization part, the text transcripts are displayed in a parallel format, a two-column view with the normalized transcript to the left and the original to the right, with highlighting in red of the deviations that have been corrected (*ний → ние, биръ → бира*), as shown in Fig. 1.

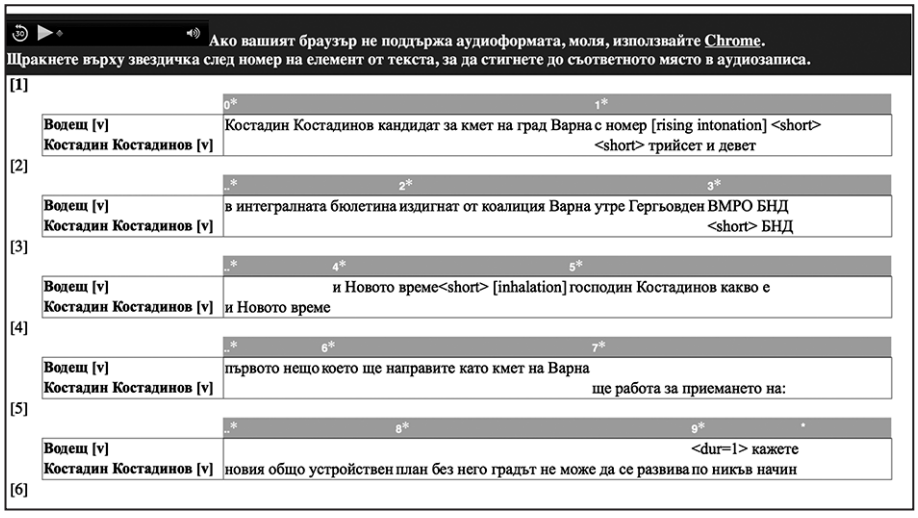
In addition, an index of this parallel corpus section was made, with a clickable list of lemmas and their attested wordforms, where a click on a wordform will lead to its instance in the text. There is also an alphabetized list of all the corrected forms, that is, all the highlighted forms in the right-side column containing the original transcript. Each form in the list is clickable and will lead to the form in its context in the original text (for more details see Dzhonova, Hauge, Tisheva 2018).

The second part of the corpus is in XML format, allowing automatic processing of the data and building displays like the one shown in Fig. 2 below, with text transcripts synchronized with sound recordings. In these transcripts as well, non-formal communication dominates. The recordings of formal communication are from media speech and interviews (Table 2).

Transcripts of spontaneous non-official communication constitute only part of the Corpus of spoken Bulgarian, due to several reasons, but mainly the need to obtain the speakers' agreement to record, transcribe, and publish. Another obstacle is the low quality of the recordings of spontaneous speech, which makes the process of transcription long and effortful. Due to this, the data col-

Table 2. Transcripts in XML format

Source	Domain	Words	Characters	Duration
audio	informal	38,491	194,346	5:13:49
audio	formal	30,539	237,415	2:46:36
TOTAL		**69,030**	**431,761**	**8:00:25**



Source: http://www.bgspeech.net/bg/resources/multimediacorpus.html

Fig. 2

lected in the last years are mostly from official communication with a low degree of preparedness. The data included in the corpus fall into several domains of communication: media speech (radio and TV recordings with more than one communicator), academic communication, and political communication. The corpus also includes a small part of business and professional communication. Such data are not easily accessible to the linguist, as they may contain confidential information. All recordings are digitalized, and in the multimedia part of the corpus the recorded sound is aligned with the textual transcript and available at a click from the user (Fig. 2).

The transcripts in the multimedia section of the corpus can be classified according to the type of media (audio or video recordings), the domain (media speech, academic, political, non-official) as follows in Table 3. Only the multimedia corpus contains transcripts of video recordings.

Table 3. Transcripts in multimedia corpus

Source	Domain	Words	Characters	Duration
video	media (TV and radio programs)	26,726	150,623	2:53:14
audio	academic	8,993	51,832	1:07:32
audio	business	38,392	297,123	3:12:05
audio	media (TV and radio programs)	47,211	390,940	4:46:45
audio	political	15,282	97,952	1:51:53
TOTAL		**136,604**	**988,470**	**13:51:29**

To sum up, the Corpus of Spoken Bulgarian consists of the following data:
- audio recordings
- video recordings
- transcripts in text format, without XML annotation, metadata included in the document
- transcripts with XML annotation of turns, speakers, pauses, overlapping, and non-verbal information
- transcripts aligned with the corresponding video or audio recording (multimedia corpus). All transcripts are available online at http://bgspeech.net/bg/resources.html.

4. Opportunities for application of the resource

Spoken corpora are increasingly used in various kinds of phonological, grammatical and lexical investigations not only of spoken language. Corpus-based studies encourage the functional, pragmatic, and communicative approaches to language investigation, teaching, and learning.

The data included in the Corpus of spoken Bulgarian enriches knowledge about the language system at all levels and provides opportunities for broadening the representativeness of linguistic research on syntax, morphology, phonology, discourse, and pragmatics. Apart from purely linguistic purposes such as verification of linguistic hypotheses, observations on linguistic dynamics and linguistic trends, spoken corpora may be useful as sources for empirical data within other fields of the humanities, such as the study of models of communication, conversational strategies (argumentation, turn-taking, expressions of bias, etc.), organization of discourse in different contexts and situations, and connections between linguistic and non-linguistic factors in communication. And apart from being a source for linguistic information, the data in spoken corpora show various successful and unsuccessful approaches in oral communication and could therefore be used in the teaching of communication strategies.

References

Chafe 1982: W. Chafe. Integration and involvement in speaking, writing, and oral literature. - In: D. Tannen (ed.). Spoken and Written Language: Exploring Orality and Literacy. Norwood: Ablex, 1982, 35-53.

Crystal 2008 (1980): D. Crystal. A Dictionary of Linguistics and Phonetics. Maxwell Publishing, Oxford, UK, 2008 (1st ed. 1980).

Dzhonova, Hauge, Tisheva 2018: M. Dzhonova, K. Hauge, Y. Tisheva. Parallel web display of transcribed spoken Bulgarian with its normalised version and an indexed list of lemmas. - In: Proceedings of the Third International Conference *Computational Linguistics in Bulgaria* (CLIB 2018), The Institute for Bulgarian Language, Bulgarian Academy of Sciences, 2018, 177-183. Available from: https://dcl.bas.bg/clib/wp-content/uploads/2018/07/CLIB_2018_Proceedings_v2_final.pdf [Accessed: 14 August 2018].

Koeva 2010: С. Коева. Българският семантично анотиран корпус - теоретични постановки. - В: С. Коева (ред.). Българският семантично анотиран корпус. София, 2010. (S. Koeva. Balgarskiyat semantichno anotiran korpus - teoretichni postanovki. - V: S. Koeva (red.). Balgarskiyat semantichno anotiran korpus. Sofia, 2010.)

Koeva, Blagoeva, Kolkovska 2010: S. Koeva, D. Blagoeva, S. Kolkovska. Bulgarian National Corpus Project. - In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valetta, ELRA, 2010, 3678-3684.

McCarthy 2004: M. McCarthy. Touchstone: From Corpus to Course Book. Cambridge University Press, 2004.

Pavlidou, Kapellidi, Karafoti 2014: T. Pavlidou, Ch. Kapellidi, E. Karafoti. The Corpus of Spoken Greek. - In: Ş. Ruhi, M. Haugh, Th. Schmidt, K. Wörner (eds.). Best Practices for Spoken Corpora in Linguistic Research. Cambridge Scholars Publishing, 2014, 56-74.

Sinclair 1996: J. Sinclair. Preliminary recommendations on Corpus Typology. Available from: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html [Accessed: 14 August 2018].

Prof. Yovka Tisheva, PhD
Faculty of Slavic Studies
Sofia University "St. Kliment Ohridski"
15 Tsar Osvoboditel blvd.
1504 Sofia, Bulgaria
Email: tisheva@uni-sofia.bg

Assoc. Prof. Marina Dzhonova, PhD
Faculty of Slavic Studies
Sofia University "St. Kliment Ohridski"
15 Tsar Osvoboditel blvd.
1504 Sofia, Bulgaria
Email: djonova@slav.uni-sofia.bg

Assoc. Prof. Kjetil Rå Hauge
Department of Literature,
Area Studies and European Languages
Faculty of Humanities
University of Oslo
0315 Oslo, Norway
Email: k.r.hauge@ilos.uio.no