THE PLACE AND ROLE OF SPEECH CORPORA IN MODERN BULGARIAN LINGUISTICS

Cvetoslava Gergovska

Abstract. *The purpose of this article is to outline the major contributions in the relatively young Bulgarian tradition in the field of corpus linguistics as well as to highlight its diversification and dynamics. Two corpora of the Bulgarian children's speech have been studied as the benefits from the research on the phenomenon of children's speech have been taken into account.*

Keywords: *corpus, corpus linguistics, electronic corpora, Bulgarian corpus linguistics, children's speech.*

INTRODUCTION

Corpus linguistics was established as a science discipline in the 50ties of the XX century. The interest in this discipline has been steadily increasing since its occurrence. Nowadays more and more linguists use corpus linguistics for their researches. "The corpus, along with the already chosen methodology and theory, becomes a mandatory part of all modern linguistic researches" [17; 3, c. 54]. Interested in corpus linguistics are not only computer specialists and experts in corpus linguistics, but also scientists in the area of lexicology, pragmatics, children's speech linguistics etc. Such an interest has resulted in the creation of different language corpora – in one language or parallel (bilingual or multilingual), written corpora or spoken corpora, united or static corpora etc.

The foreign scientific practice had proved the benefit of the corpora and the speech researches. Suffice it to mention the names of McEnery, Wilson, Biber, Leech and others. In this paper an attempt will be made to present the development of the Bulgarian corpus linguistics and its unique path from the days of its origin to modern times, which follows the domestic traditions in linguistics, on the one hand, and the contemporary linguistics as a whole on the other. Children's speech corpora shall be presented as well, in order to clarify the role of the corpus in the speech studying process.

Before proceeding with the review of the contributions in Bulgarian corpus linguistics we need to clarify what is meant by the term "corpus". Numerous definitions exist in the world linguistic literature explaining the essence of the corpus. One of the most popular definitions is provided by McEnery and Wilson. According to them, the corpus is **«**a finite collection of machine-readable, sampled to be maximally representative of a language or variety**»** [10, c. 197]. S. Koeva gives the following definition

within the scope of Bulgarian linguistics: «A corpus is a large collection of language samples, presented in a way suitable for computer processing and chosen according to perparticular (linguistic) criteria, so they represent an adequate language model» [6, c. 9].

Y. Tisheva and M. Dzhonova define the corpus as a «collection (an aggregate, a system organization) of texts in an electronic format, that have gone through different stages of reprocessing (annotated) according to different linguistic and extra linguistic parameters» [17; 3, c. 55].

The present article shall consider the term corpus as a collection of systemized texts electronically stored and used for linguistic purposes.

SPEECH CORPORA IN MODERN BULGARIAN LINGUISTICS

The present publication does not claim to be an exhaustive presentation of all existing corpora in Bulgaria and it is only directed towards the main highlights. Most popular contributors in that direction are: S. Koeva, D. Blagoeva, S. Kolkovska, Y. Tisheva, M. Dzhonova, K. Simova, P. Osenova, V. Popova, Ts. Nikolova, and Ts. Venkova. The Bulgarian National Corpus, The Corpus of spoken Bulgarian of the BgSpeech Initiative, Nikolova-Venkova Corpus and the Corpus of the Bulgarian Children's Speech, developed by the Laboratory in Applied Linguistics of the University of Shumen are amongst the most prominent corpora.

In 2009 some researches from the Section in Computer Linguistics and the Section for Bulgarian Lexicology and Lexicography of "Prof. L. Andreychin" Institute for Bulgarian Language, part of the Bulgarian Academy of Science established the Bulgarian National Corpus [18]. This corpus unites a few separate electronic corpora that had been developed in the period from 2001 to 2009 for the use of those two Sections. The Bulgarian National Corpus (BNC) consists of a Bulgarian part and 47 other parallel corpora. The first part consists of approximately 1.2 billion words and includes more than 240 000 texts. The materials present the Bulgarian language in the period from 1945 until now. Most of the materials examine the written form of the Bulgarian language. BNC is a convenient source to observe the frequency of usage of words or language structures, generating frequency lists etc. It is possible to search for samples of particular linguistic occurrence in order to obtain a linguistic description, lexicographical incision, or for study purposes.

The corpus with spoken Bulgarian is more closely specialized and it is established by the creators of BgSpeech project [19]. The corpus consists of 4 parts: transcribed spoken Bulgarian (2001-2004), transcribed spoken Bulgarian (2004), multimedia corpus, and a parallel corpus. The transcriptions are of use for different scientific and educational purposes. Publications on issues of the colloquial language and Bulgarian dialects are also accessible on the site of BgSpeech project.

Another corpus of spoken Bulgarian that needs to be highlighted is the "Nikolova-Venkova" corpus. As Ts. Venkova says, the corpus is defined as one of the harbingers of the modern electronic corpora. The

"Nikolova-Venkova" corpus is not large in volume (approximately 50 000 word forms). It is constructed of texts with no annotation. The corpus of Bulgarian verbal speech is formed on two stages. The first stage was the collection of a massif of typewritten texts that are transcriptions of tape-recorded conversations. It was realized by Tsvetanka Nikolova in the period from 1975 to 1979 as the purpose was the creation of a Frequency dictionary of Colloquial Bulgarian, published in 1987. Ts. Venkova transcribes the recordings and creates an Electronic corpus of the conversational speech (1993 – 1995) in the second stage. The corpus is available at [20].

The other corpus to be presented herein is the Syntactic Bank of the Bulgarian language (BulTreeBank). The BulTreeBank [21] consists of approximately 15 000 sentences, that may be used for automatic extraction of linguistic knowledge – taggers, dictionary information, parser. The BulTreeBank Group is working on projects related to Computational Linguistics and Semantic Web. Their primary purpose is to create Language Resources and Tools for Bulgarian. They have worked on creation of a Bulgarian Treebank, POS tagger, Partial Grammar, Text Archive, Domain Ontologies, Lexicons, XML Tools. The BulTreeBank Group is part of the Linguistic Modelling Laboratory (LML), Institute of Information and Communication Technologies , Bulgarian Academy of Sciences.

The electronic corpora are a labour-intensive and sometimes expensive undertaking, but at the same time they are extremely necessary in cases when specific speech phenomena are being examined; children's speech is amongst them. The interest towards the speech of the children never subsides and interests many researchers nowadays as well. The children's speech is a vast area, giving important information related to the solution of numerous theoretical problems related to the mechanisms of learning and usage of a language in the process of speech communication. All processes related to the formation of language can be observed through the means of corpora with Bulgarian children speech. Such corpora give the opportunity to observe the speech of children of the same or different age. Another advantage is that by comparing the language data of children and adults one can follow the speech evolution, the development stages (starting from one-word statements, through the two-word statements and polyword statements, reaching to a correctly structured speech).

All of the aforesaid predetermines the interest of the team of the Laboratory in Applied Linguistics (**LabLing**) in the University of Shumen towards the inclusion of the **Electronic corpus in Bulgarian Children's Speech** in the work program for creation of corpora with a spontaneous speech, is a result of long years of work which had started in 1990 and still in progress.The initiators of this project are the Director of **LabLing** Dimitar Popov and the researcher Velka Popova.

The electronic corpus of Bulgarian children's speech is processed in CHILDES system's terms [22]. The choice is motivated by its author Velka Popova as an extremely useful and suitable platform for research work because of the typology diversity and the integrated speech data, unified transcription format and the package of program resources CLAN for automatic reprocessing [see: 16]. The favorable

empirical opportunities provide for the linguistic research such characteristics as neutrality and adequacy of the results and give a «solid base for approbation of the speech ontogenesis models» **[14, c. 294]**. CHILDES system is extremely necessary and useful when realizing large integrative examinations of children's speech within the scope of international scientific projects (for example – cross-linguistic projects Pre- and Protomorphology in Language Acquisition; Syntaktische Konsequenzen des Morphologieerwerbs; Erwerb sprachlicher Markierungen zur Differenzierung von ±Begrenztheit; Spracherwerb: Acquisition and Disambiguation of Intersentential Pronominal Reference etc. – [see: **14, c. 292**].

The corpus consists of two types of speech resources – sub-corpus A - spontaneous speech of 4 children in early age (from 1 to 3 years old) and a sub-corpus B, consisting of tales based on series of pictures told by 90 children in a preschool age (from 3 to 6 years old). The audio recordings are transcribed and coded in CHAT-format. This is a way for the package to be used with CLAN specialized programs and gives opportunity to analyze the dialogues and their comments. Another advantage is that each user is allowed to create additional comments depending on the purposes of the particular examination. «The information, brought by the additional comments is particularly important at examinations both of children's speech as it is full of deviations and is strongly-dependent on the situation and for learning second language (L2), interaction between adults and children as well as for the usage of more CLAN programs, which makes them useful for a wide scope of specialists.» [**16**] The feasibility of the corpus with speech data of Bulgarian children is partially approbated within the frame of discussions and comparative analysis of the Bulgarian with other languages (German and Russian), made within the cross-linguistic program for examination of the aspect early assimilation. (comp. **[8; 2]**. The corpus is an empiric base for numerous private examinations onto different aspects of the early ontogenesis of the Bulgarian grammar [**11; 12; 13 etc.**] The electronic corpus of Bulgarian children's speech is included in the program for design of BG-CLARIN, a part of CLARIN CHILDES (Common Language Resources and Technology Infrastructure).

A module, developed by the author of the present article is a stage in the creation of the electronic corpus for Bulgarian children's speech, also created in CHILDES. This sub-corpus is still in a working regime (the planning stage and the empiric examinations already took place as well as the digital recordings and the data transcription – entirely done by the author of the article herein) and it is going to be structured in two parts.

The first part consists of 18 hours of children's speech recordings. The corpus is created in connection with a research on syntactic synonymy in the modern Bulgarian children's speech. The recordings were done in two kindergartens in Shumen. All children were of preschool age (from 3 to 6 years old) to the total number of 27. The children – participants live in Shumen, Northeastern Bulgaria. Some of the children are bilingual (Bulgarian and Turkish language). Their parents are of good level of education –

high school or university. The corpus is presented in 27 files in CHAT-format (the speech of each child in a separate file). The recordings were done one by one. Three tales in pictures were presented to the children. Then the researcher told each tale prepared text. After a short conversation the child is supposed to reproduce the tales. For illustration you may see Figure 1:

Figure 1. Transcript

@Begin
@Participants: Ber Target_Child, CVE INVESTIGATOR
@Birth of BER: 2009
@Age of BER: 6

*CVE: Az se kazvam Cveti.
*CVE: Ti kak se kazvash?
* BER: Az se kazvam Beren.
*CVE: Na kolko godinki si?
* BER: Na shes.
*CVE: Na shes!
……

*CVE: Tochno kogato Sharo doprql nosleto si do cveteto, ot cveteto izlqzla edna pchelichka.
* BER: Tykmo da si podade nosleto i ot cveteto izlqzla pchela.
……

*CVE: Pilenceto se natyzhilo i dobavilo:
*CVE: Az nqma da dojda, ponezhe ne moga da pluvam.
*BER: Pilenceto mu otgovorilo:
* BER: Ne moga, az ne moga da pluvam.
……...
@End

The second part of the corpus consists of 15 hours of speech recording of 5 children (one of the children is 3 years old and the others – 6 years old). The recordings here are presented in 5 CHAT-files (each speech in a separate file). The children's speech had been recorded in different time. The recordings

took place while the children are being dressed, falling asleep, eating or playing games. All children are monolingual. Their parents are of good level of education – high school or university. The children come from different cities – Sofia (West Bulgaria), Montana (North-west Bulgaria) and Knezha (Central North Bulgaria).

Conclusion

As it was mentioned above the purpose of the present article is to present the major accents of the most popular Bulgarian corpora. This short review shows that the already existing extensive databases are in a process of continuous development and growth and their inclusion in different international projects may be interpreted as an ambiguous guarantee for their reliability and utility.

We may conclude that corpus lingustics is widely established as a preferred research platform for the Bulgarian scientists who do research in the field of speech.

References

1. Biber et al. Corpus Linguistics: Investigating Language Structure and Use. - Cambridge, Cambridge University Press, 1998.
2. Bitner, D., Gagarina, N., Popova, V. Kühnast, M. Aspect before Tense in the acquisition of Russian, Bulgarian, and German. // V. Solovyev, V. Polyakov (Eds.). Text Processing and cognitive Technologies. - Moskow, 2005.
3. Dzhonova, M. Corpus with Bulgarian verbal speech - specificity and structure. // Bulgarian language, 2011. - № 58. - P. 34 – 53. (In Bulgarian)
4. Koeva, S. Language resources and computer programs with application in linguistic research. // IT guide to humanitarians [Internet]. - Plovdiv: Plovdiv university „Paisii Hilendarski“, 2009. - P. 30–53. Available from: http://dcl.bas.bg/PDF/LanguageResources.pdf. (In Bulgarian)
5. Koeva, S., Stoqnova, I. Bulgarian National Corpus. // Bulgarian language. - 2009. - P. 135–150. (In Bulgarian)
6. Koeva, S., Blagoeva, D., Kolkovska, S. Bulgarian language electronic corpus and its applications. // Journal Science, 2010. - № 5. - P. 64-69. (In Bulgarian)
7. Koeva, S., Blagoeva, D., Kolkovska, S. The project Bulgarian national corpus – results and prospects. // Bulgarian language, 2011. - № 3. - P. 34-53. (In Bulgarian)
8. Kyunast, M., Popova, V., Popov, D. Erwerb der Aspektmarkierung im Bulgarischen. // N. Gagarina, D. Bittner (Eds.). ZAS-Paper in Linguistics 33, 2004: Studies on the development of grammar in German, Russian and Bulgarian, 2004. - P. 63-87. Available from http://alphalinguistica.sns.it/Riviste/ZAS/ 33_2004.pdf.

9. Leech, G. Introducing corpus annotation. – In: R. Garside, G. Leech, A. M. MacEnery (eds.). Corpus Annotation: Linguistics Information from Computer Text Corpora. - London, Longman, 1997. - P. 1-18.

10. McEnery, T., A. Wilson. Corpus Linguistics. - Edinburgh, Edinburgh University Press, 2001.

11. Popova, V., Popov, D. The emergence of verb grammar in two Bulgarian-speaking children. // V. Solovyev, V. Polyakov (Eds.). Text Processing and cognitive Technologies. - Moskow, 2007. - P. 236-248.

12. Popova, V. Corpus research of grammatical metamorphosis of early child language**.** // Language, culture, identity. - Veliko Tyrnovo: Faber, 2010. - P. 101-115. (In Bulgarian)

13. Popova, V. The role of onomatopeia in early verb ontogenesis – Litera et Lingua. Spring. - Sofia, 2011. Available from http://slav.unisofia.bg/lilijournal/index.php/ bg/issues/spring2011. (In Bulgarian)

14. Popova, V. Corpus perspective in the study of children's speech. //40 years Shumen university – 1971-2011. - 2012. - P. 286-294. (In Bulgarian)

15. Popova, V., Popov. D. Investigating speech interaction from holistic integrative point of view **.** // S. Jovičić, M. Subotić, M. Sovilj (Eds.). VERBAL COMMUNICATION QUALITY. Interdisciplinary Research II. - Belgrade: LAAC & IEPSP, 2013. - P**.** 373-396**.**

16. Popova, V. Electronic corpus of Bulgarian children's speech. In: Paisievi cheteniq. - Plovdiv university „Paisii Hilendarski", 2014. (Forthcoming) (In Bulgarian)

17. Tisheva, Y. Language databases, corpora and electronic resources for Bulgarian spoken speech. Available from http://slav.uni-sofia.bg/naum/lilijournal/

18. 2014/11/1-2/ytisheva. (In Bulgarian)

19. http://dcl.bas.bg/bulnc/

20. http://www.bgspeech.net/index.html

21. www.t.venkova.info.

22. http://www.bultreebank.org/

23. http://childes.psy.cmu.ed).

Authors' Information

Cvetoslava Gergovska Konstantin Preslavski University of Shumen. Shumen. Bulgaria.
e-mail: cve_lazarova@abv.bg