https://doi.org/10.33919/JCAL.23.1.3

COMPUTER-ASSISTED TRANSCRIPTION AND ANALYSIS OF BULGARIAN CHILD SPEECH DATA USING CHILDES AND CLAN

Velka Popova[1], Dimitar Popov[1]

[1] *Konstantin Preslavsky University of Shumen*

Abstract

The present paper focuses on the possibilities offered by corpus linguistics in the study of child speech, with its specificities as a linguistic phenomenon. An attempt is made to highlight the advantages of the CHILDES system for studying spontaneous speech interaction in the Bulgarian corpus of child language (Bulgarian LabLing Corpus), in which the data are transcribed and annotated within this paradigm.

Keywords: *CHILDES, CLAN, Bulgarian LabLing Corpus*

1. Introduction

In recent decades, linguistic resources organised as corpora, have been increasingly used in the modelling of language and speech behaviour of its speakers, despite the fact that the creation and maintenance of computerised corpora is extremely laborious and costly. Modern technologies required a new, more effective standard for data presentation and processing. They made it possible to extend the scope of a corpus to millions of language items while also optimising the options for their annotation (linguistic analysis), unification, standardisation and repeated use. The magnitude of change in research is even more evident in the application of modern computer programmes for automatic processing of huge databases in the corpus approach to research in the field of humanities.

The present paper focuses on the possibilities offered by corpus linguistics in the study of child speech, with its specificities as a linguistic phenomenon. An attempt is made to highlight the advantages of the CHILDES system for studying spontaneous speech interaction in the Bulgarian corpus of child language (Bulgarian LabLing Corpus), in which the data are transcribed and annotated within this paradigm. The corpus is available to researchers at:

https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html.

2. Corpus paradigm – necessity or fad in child language research

The proposed study addresses the question of whether the use of a corpus paradigm in studying child language is a necessity or a fad, having in mind that the creation of computerised corpora is an extremely costly and laborious endeavour. There is the question if child language as a specific linguistic phenomenon is worthy of studying by means of sophisticated research tools. This in turn leads to the question whether we need to study child language at all, and what the advantages of the corpus paradigm are in comparison to some traditional approaches which have been used so far in the study of linguistic ontogenesis.

In the linguistics tradition there has been an enduring interest in the phenomenon of child language and this is not only for the sake of studying it or out of mere research curiosity. On the contrary, data about linguistic ontogenesis are in many cases the mandatory ‘external evidence’ for testing different hypotheses or theoretical constructs. Along with their importance for clarifying issues in linguistic typology and universals, these data are crucial in resolving problems of early speech pathology and language teaching. In recent decades, in line with the fast developments in psycholinguistics and cognitive linguistics, child language has also proved to be the key to the hidden functioning of the human perceptive and cognitive faculties. In this way, the research of linguistic ontogenesis is part of the general tendency in modern linguistics to overcome isolated study of language for the sake of it and focus on human speech interaction.

The importance of child language necessitates the creation of an adequate model of language ontogenesis, which in turn raises the question of the suitability and (in)sufficiency of the approaches which can help debunk myths not only in linguistics, but also in psycholinguistics, and in the theory of language learning. This in turn requires the use of relevant models and reliable empirical material.

Next comes the question of the methods for collecting sufficient quantitative and qualitative empirical evidence for adequate study of child language. In this regard, it should be noted that ever since Charles Darwin, the corpus approach has been a major factor in the research on language ontogenesis. There is ample evidence in support of such an approach.

Developments in technology over time have brought about a new quality of empirical data and the possibilities for their processing. Files and diaries have been replaced with electronic data of speech recordings, and the hard, intensive and exhausting work involved in the registration, transcription and statistical processing of data is now replaced by modern technology and software products. The apogee of this evolutionary process is the creation of the CHILDES system. The typological diversity of the included linguistic data, the unified manner of transcription, and the package of programme resources CLAN for automatic processing, turn this system into an extremely useful and convenient research platform. In the following section of this paper we will use the example of the Bulgarian CHILDES-Corpus to demonstrate its tools for empirical verification of the models of language ontogenesis. We will look for an answer to the question of the role of the computerised CHILDES system in overcoming difficulties associated with the specifics of child language.

3. Bulgarian resources of spontaneous child speech in CHILDES terminology

In the autumn of 2020, a new addition appeared in the database of the CHILDES platform in its Slavic languages section, namely the *Bulgarian LabLing Corpus*. It appeared as a result of long years of work done by researchers from LABLING. The corpus data are transcribed in the unified CHAT format of the CHILDES system (MacWhinney, 2010), which makes them comparable to the corpora in other languages in the platform. Long before the publication of the Bulgarian LabLing Corpus, the application and reliability of this base comprising speech data and information obtained from Bulgarian children was partially approbated in discussions and comparative analyses of Bulgarian and the other languages (in particular, German and

Russian), carried out in the sphere of cross-linguistic programme for examining the early adoption and mastering of the aspect (comp.: Kühnast et al., 2004; Bittner et al., 2005). The corpus also stresses the empirical base and the multiplicity of particular research works on different aspects of the early-age ontogenesis of Bulgarian grammar. Bulgarian computarised empirical data have been used in the process of the empirical verification of the pre- and proto-morphology model (see: Popova, 2007; Popova, 2016).

The present study is based on the longitudinal subcorpus of the Bulgarian LabLing corpus. This collection comprises spontaneous speech samples produced by five children aged between 1 and 3 years. In the core of the database there are 47 hours of recordings. They are presented on the CHILDES platform in 104 files in CHAT-format.

4. How does CHILDES provide sufficient and reliable information needed for linguistic analysis?

The Bulgarian child language corpus was created by using the two main tools of the CHILDES system: the special format for transcribing and coding - CHAT, and the package of programmes for analysis - CLAN.

The file format in CHILDES is called CHAT and all files are saved as *.cha. For transcription and playback, the relevant part of CLAN is the editor. The editor uses most of the same conventions as Microsoft Word. However, unlike Word, it allows the researcher to link individual segments of the transcript directly to the audio or video media.

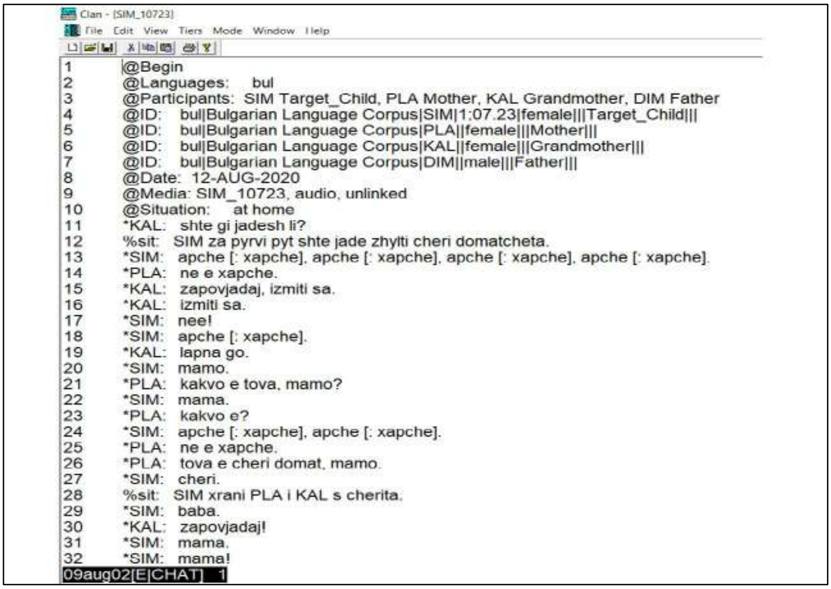
Within CHILDES the Bulgarian resources of spontaneous child speech are presented in the mandatory CHAT format (MacWhinney, 2000). It includes the following: 1. Title lines, containing information about the participants in the dialogue, their age, date of birth, date and conditions of the recording; 2. The alternating utterances of the participants, formed as single lines and the accompanying comments, given as additional lines. The headings were chosen by the researcher, and @Begin, @Participants and @End were left as mandatory, as in the following sample transcript:

```
@Begin
@Participants:      ALE Alexandra Target_Child, VEL Velka Mother
@ALE's birthdate:  29-JAN-1989
@Date:      27-MAR-1990
@Filename:          al10129
@Age of ALE:        1;01.29
@Situation:         at home
```

```
*VEL:    [spoken material]
*ALE:    [spoken material]
*VEL:    [spoken material]
@End
```

For modern studies of early linguistic ontogenesis in the context of constructivism, it is particularly important that the data not only of children but also of adults who care for them is taken into consideration. With the CHILDES standard for transcription, optimal conditions are created for them to be adequately described, as it includes mandatory order Participants. Here, together with their names and social roles, a three-letter code is introduced, which starts the lines of each participant in the subsequent dialogue. Thus, optimal conditions are created to isolate, monitor and analyse the lines of each participant, which would lead to a more objective study of child speech and child-directed speech.

In this regard, a wide applicability of the Longitudinal Corpus of the Bulgarian CHILDES collection can be expected, as each of the transcripts includes data on the identification of the participants (demographic and linguistic parameters) and the respective corpus. See Fig. 1:



```
@Begin
@Languages:   bul
@Participants:  SIM Target_Child, PLA Mother, KAL Grandmother, DIM Father
@ID:   bul|Bulgarian Language Corpus|SIM|1;07.23|female|||Target_Child|||
@ID:   bul|Bulgarian Language Corpus|PLA||female|||Mother|||
@ID:   bul|Bulgarian Language Corpus|KAL||female|||Grandmother|||
@ID:   bul|Bulgarian Language Corpus|DIM||male|||Father|||
@Date:  12-AUG-2020
@Media: SIM_10723, audio, unlinked
@Situation:    at home
*KAL:  shte gi jadesh li?
%sit:  SIM za pyrvi pyt shte jade zhylti cheri domatcheta.
*SIM:  apche [: xapche], apche [: xapche], apche [: xapche], apche [: xapche].
*PLA:  ne e xapche.
*KAL:  zapovjadaj, izmiti sa.
*KAL:  izmiti sa.
*SIM:  nee!
*SIM:  apche [: xapche].
*KAL:  lapna go.
*SIM:  mamo.
*PLA:  kakvo e tova, mamo?
*SIM:  mama.
*PLA:  kakvo e?
*SIM:  apche [: xapche], apche [: xapche].
*PLA:  ne e xapche.
*PLA:  tova e cheri domat, mamo.
*SIM:  cheri.
%sit:  SIM xrani PLA i KAL s cherita.
*SIM:  baba.
*KAL:  zapovjadaj!
*SIM:  mama.
*SIM:  mama!
```

Fig. 1: A fragment of a transcript from the Bulgarian LabLing Corpus

Another important advantage of the CHAT-form of the transcripts is that, in view of the specific objectives that the researcher has in a given study, he or she can add lines of comments as needed. The comments can be of different nature: phonetic, morphological, situational, by the author, respectively, presented in the CHAT file as special lines: %Pho, %mor, % sit, %com, etc. For example, in the Bulgarian CHILDES-corpus additional %sit and %com lines are introduced, as well as the short presentation of deviations from the target language norm are given immediately before [: the norm unit]. Organised in this way, the speech resources prove to be very important as a reliable empirical basis for studying children's speech, as it is highly situational, abounds in deviations from the norm, and the values of the lexical deficit are too high. The following fragments of the Bulgarian CHAT-transcripts illustrate well these points:

TEF, 1;11:
*TEF: Nyama dam!
"No, I'm not giving it to you!"
%sit: TEF jumps on the bed and BAB is trying to catch her hand so that she won't fall. TEF keeps on jumping and puts her hands behind her back.
(2) ALE (1;1)
*ALE: Mama, mama!
"Mummy, mummy!"
%sit: She points to the door.
*VEL: Pri mama li iskaš?
"You want [to come] to mummy?"
%sit: VEL provokes the child by pretending not to understand the child's message.
*ALE: Mamo, mamo!
"Mummy, mummy!"
%sit: She implores with a whining voice.
*ALE: Mamma, mamma, maamma!
%sit: ALE pulls VEL rudely to the door.

The CHILDES provides the researchers with a package of specialised CLAN programmemes, which on their part can implement different types of analysis of the inserted dialogues. In that respect CLAN can automatically provide diverse statistical and substantial results out of the transcribed and coded data such as word frequency, lexical diversity and combinations, about a specific user's words and forms (for example, child language errors, such as specific deviations from the norm of the given language: the units of the so-called Baby Talk, onomatopoeia, super-generalisations, child and family occasions), etc. CLAN consists of two parts, namely the editor and the programmes. The second part of the CLAN provides the programmes for searching and analysis. The CLAN programmes which are of great interest with respect to interaction analysis include CHIP, COMBO, GEM, KWAL, and TIMEDUR. COMBO and KWAL allow users to search for all types of word and symbol combinations (MacWhinney and Wagner, 2010).

As an illustration to the aforementioned (see below) we will turn to the corpus of a Bulgarian girl – Alexandra (marked in the transcription with ALE) in order to demonstrate how conveniently and fast via FREQ programmeme from the CLAN set a frequent analysis could be implemented regarding the coded onomatopoeic elements in the main lines of the investigated child. After the start of CLAN, first we open the file (namely <probe.cha>).

Initial file: <probe.cha>

```
@Begin
@Participants:      ALE Alexandra Target_Child, VEL Velka Mother
@Birth of ALE:      29-JAN-1989
@Date:        27-MAR-1990
@Filename:          probe.cha
@Age of ALE:        1;01.29
@Situation:         at home
*ALE: Pyche [:pypche].
%sit:         poglezhda pypcheto si
*VEL:         Kyde e guceto grux-grux?
*ALE:         Gux-gux@o.
*VEL:         Grux-grux?
*VEL:         A kucheto, mamo, kyde e?
*VEL:         Kuche-e!
*VEL:         Au, njama go kucheto!
*VEL:         Kyde e kucheto?
*VEL:         am?
*ALE:         Bau-bau@o!
*VEL:         Vizh kakvo dyrzhi tati?
%sit:         pokazva kartinka s momiche, dyrzhashto kuchence.
*ALE:         Bau-bau@o.
@End
```

After that we press the command icon Commands, in which the necessary formula is typed (namely – freq +t*ALE +k +d*@o* +f probe.cha). Then we activate the operation and if the control system does not find any errors in the document structure, a new file is created (in particular – <probe.fr0.cex>), containing a list of child utterances with onomatopoeic elements (the trajectories of use of which, ordered alphabetically, are signed in the original *.cha file) and a quantitative analysis of the frequency of the coded elements.

```
Final file 1 (OUTPUT 2): <probe.fr0.cex>
freq +t*ALE +k +d*@o* +f probe.cha
Fri May 27 04:24:20 2005
freq (13-Apr-2001) is conducting analyses on:
ONLY speaker main tiers matching: *ALE
****************************************
From file <probe.cha> to file <probe.fr0.cex>
2 Bau-bau@o     : 19,22
1 Gux-gux@o     : 12
1 Pyche         : 9
------------------------------
3  Total number of different word types used
4  Total number of words (tokens)
0.750  Type/Token ratio
```

Thanks to the resources of the programme package CLAN, from files *.cha at the exit, it is possible to obtain different types of files. They could give information, which is necessary not only for statistical, but also for meaningful analysis of the corresponding chunk of speech. Particularly useful in this respect is the possibility (which is given by the KWAL command) to obtain an exit file (see below <probe.kw0.cex>) with isolated rows containing the element the researcher is interested in with the exact trajectories marked. In this way, the command Go in the source text can interpret the conversion context immediately.

```
Final file 2 (OUTPUT 2): <probe.kw0.cex>
kwal +t*ALE +k +s*@o* +f probe.cha
Wed Jun 01 10:21:24 2005
kwal (13-Apr-2001) is conducting analyses on:
 ONLY speaker main tiers matching: *ALE
****************************************
From file <probe.cha> to file <probe.kw0.cex>
----------------------------------------
*** File "probe.cha": line 12. Keyword: @o
*ALE:       Gux-gux @o.
```

```
--------------------------------------
*** File “probe.cha”: line 19. Keyword: @o
*ALE:        Bau-bau @o!
--------------------------------------
*** File “probe.cha”: line 22. Keyword: @o
*ALE:        Bau-bau @o.
```

Another advantage of the programme is that the main lines of the investigated child can be isolated not only independently but also in the context of one or several preceding or following lines, which is of particular importance for the research of speech ontogenesis. For example, from the initial demonstration file <probe.cha> we can receive an exit file containing information about the context of the child’s utterance (in this case it is specified as a necessary line preceding the child’s speech, which is encoded in the exit file with [– W1], resulting in the <probe.kw1.cex> file.

```
Final file 3 (OUTPUT 3): <probe.kw1.cex>
kwal +t*ALE +k +s*@o* -w1 +f probe.cha
Sun Jun 05 10:53:29 2005
kwal (13-Apr-2001) is conducting analyses on:
  ONLY speaker main tiers matching: *ALE
**************************************
From file <probe.cha> to file <probe.kw1.cex>
--------------------------------------
*** File “probe.cha”: line 12. Keyword: @o
*VEL:        Kyde e guceto grux-grux?
*ALE:        Gux-gux @o.
--------------------------------------
*** File “probe.cha”: line 19. Keyword: @o
*VEL:        Tam ?
*ALE:        Bau-bau @o!
--------------------------------------
*** File «probe.cha»: line 22. Keyword: @o
*VEL:        Vizh kakvo dyrzhi tati?
*ALE:        Bau-bau @o.
```

The contextual presentation of the analysed linguistic phenomena plays a vital role in the study of language categories, which are associated with sophisticated semantic complexes, as the preliminary treatment of the corpus prevents potential problems caused by polysemy and homonymy.

5. Conclusion

The sample demonstrations presented here do not exhaust all the advantages of the CHILDES system in the study of linguistic ontogenesis. The easy and user-friendly procedure for quantitative analysis, as well as the fact that CLAN is constantly improving, characterise it as a dynamic, efficient and convenient programme for working with large speech databases. This is what determines the widespread use of CLAN resources in the processing of the empirical material underlying modern models of linguistic ontogenesis.

The importance of CHILDES corpora and CLAN package of computer programmes can be summarised in the following potential applications:

Explanation in parent-child conversation using the CHILDES database;

- Modelling of input language system;
- Adequate and economical presentation of data related to highly deviant spontaneous child speech;
- Concise yet sufficient presentation of extralinguistic information, needed for the understanding of children's utterances;
- Automatic processing of speech databases;
- Automatic quantitative and statistical data analysis;
- Solving problems arising from homonymy, polysemy and other linguistic phenomena;
- Linguistic analysis at different levels;
- Multi-modality interface which allows for repeated use (see Popov and Popova, 2015).

In conclusion, it should be noted that the publication of the Bulgarian LabLing Corpus in the CHILDES system leads to an expansion of cross-linguistic research by adding another Slavic language to the database. In addition, the Bulgarian linguistic tradition acquires another universal easy-to-use standard for studying linguistic ontogenesis, thanks to which scientists will have the opportunity to quickly, accurately and reliably make comparisons among a large number of languages and build adequate typologies and sound modern theories.

6. Acknowledgements

This research was partially funded by the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favour of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG, Grant number DO1-272/16.12.2019.

References

MacWhinney, B. (2000). *The CHILDES Project. Tools for Analyzing Talk. Vol. II, The Database*. Hillsdale: Lawrence Erlbaum Associates.

MacWhinney, B. and Wagner, J. (2010). Transcribing, Searching and Data Sharing: The CLAN Software and the TalkBank Data Repository. Gesprächsforschung. *Online-Zeitschrift zur verbalen Interaktion,* 11, 154–173. Available at: www.gespraechsforschung-ozs.de.

Bittner, D. et al. (2005). Aspect Before Tense in the Acquisition of Russian, Bulgarian, and German. In: *Text Processing and cognitive Technologies. V. 2. Proceedings of the VIII-th International Conference Cognitive Modeling in Linguistics (CML 2005)*. Moscow: MISA, Ucheba, 263–272.

Popov, D. and Popova, V. (2015). Multimodal Presentation of Bulgarian Child Language. In: *Proceedings of the 17th International Conference Speech and Computer (SPECOM 2015).* Springer International Publishing Switzerland, 293–300. Available at: https://www.springer.com/gp/book/9783319231310.

Kühnast, M. et al. (2004). Erwerb der Aspektmarkierung im Bulgarischen. *ZAS-Paper in Linguistics (Studies on the development of grammar in German, Russian and Bulgarian),* 33, 63–87. Available at: https://d-nb.info/1054690154/34.

Popova, V. (2006). *Child Language Early Grammar. Cognitive Aspects of Verbal Ontogenesis.* Veliko Tarnovo: Faber. (In Bulgarian)

Popova, V. (2016). *Event Modality. Early Ontogenesis*. Shumen: Konstantin Preslavsky University Publishing House. (In Bulgarian)