LABLASS and the BULGARIAN LABLING CORPUS for Teaching Linguistics

Velka Popova
Laboratory of Applied Linguistics, Konstantin Preslavsky University of Shumen, Bulgaria
v.popova@shu.bg

Radostina Iglikova
Laboratory of Applied Linguistics, Konstantin Preslavsky University of Shumen, Bulgaria
r.iglikova@shu.bg

Krasimir Kordov
Laboratory of Applied Linguistics, Konstantin Preslavsky University of Shumen, Bulgaria
krasimir.kordov@shu.bg

Abstract

The article reviews the first steps in integrating CLARIN into the curriculum at Konstantin Preslavsky University in Shumen, Bulgaria. It discusses the transition from informational seminars for undergraduate and PhD students of different majors regarding the possibilities of this European interdisciplinary network all the way to the specific first steps towards integrating its resources and instruments in the process of education. The focus here is on the approbation of resources and instruments developed within the ClaDA-BG project and, specifically, on the application of two products of the LABLING Laboratory of Applied Linguistics at Shumen University as technological partner to the National Consortium ClaDA-BG - the LABLASS web-based system for researching free speech associations and the BULGARIAN LABLING CORPUS of systematized child speech data. The introduction of their pilot versions emphasizes their importance for achieving higher standards of research work although the accent falls on their application within teaching Linguistics. They concern the updating of the curriculum content and the practical modules of various linguistic disciplines, the creation of new resources, the introduction of interdisciplinary scenarios for classwork with teachers of different academic backgrounds (a linguist and an IT specialist), as well as the transitioning of the teaching process out of the classroom and into the research lab.

1 Introduction

The CLaDA-BG national infrastructure aims at supplying resources and instruments to arts-, humanities-, and social studies researchers with the expectation that the actual applications will exceed the research frame and the results will thus be applicable to the field of education as well (Osenova and Simov, 2020). This idea fits naturally in the new *CLARIN in the Classroom* initiative which has to do with the inclusion of CLARIN resources, instruments and services in university education where the competences paradigm is continuously being recognized as one of the main ways of resolving the crisis between the accumulation of large volumes of knowledge and the inability of the students to use it in practice (for more details see (Popova, 2018)). And since the research process guided by the idea of "learning through exploration" to a large extent determines the competences paradigm in education, the *CLARIN in the Classroom* initiative can be recognized as timely and useful.

The article reviews the first steps in integrating CLARIN into the curriculum at Shumen University. The transition from informational seminars for undergraduate and PhD students of different majors regarding the possibilities of this European interdisciplinary network all the way to the specific first steps towards integrating its resources and instruments in education is discussed. The focus here is on the approbation of resources and instruments developed within the ClaDa-BG project and specifically on the application of two products of the LABLING Laboratory of Applied Linguistics at Shumen University as technological partner to the National Consortium ClaDa-BG – the LABLASS web-based system for researching free speech

Velka Popova, Radostina Iglikova and Krasimir Kordov 2021. LABLASS and the BULGARIAN LABLING CORPUS for Teaching Linguistics. *Selected papers from the CLARIN Annual Conference 2020.* Linköping Electronic Conference Proceedings 180: 180 208–213.

associations and the BULGARIAN LABLING CORPUS of systematized child speech data. The introduction of their pilot versions emphasizes their importance for achieving higher standards of research work, although the accent falls on their application within the teaching of Linguistics.

2 The Role of the General Information and Specialized Seminars for the Professional Development of the Student-Participants in the CLaDA-BG Project

The *CLARIN in the Classroom* initiative would not have found fertile ground for development in the education process at Shumen University if it had not been preceded by informational events which provided a high level of awareness among the academic staff regarding their wide applicability in the social sciences and the humanities. Therefore, immediately after Bulgaria signed the Memorandum of Understanding in 2012, thus becoming one of the 9 founding members of CLARIN ERIC, LABLING began to organize and hold annual seminars for undergraduate and PhD students from different majors to popularize the already available resources and functionalities of the network for interdisciplinary interaction, having discussions regarding the problems of corpus linguistics and the creation of digital linguistic models.

The seminars were organized in the context of excellent cooperation and mutual assistance between the separate partner associations within CLaDA-BG. This ensured that undergraduate and PhD students had access to information regarding the resources, instruments and services contributed not only by researchers within LABLING but also by other teams within the Bulgarian consortium and the European CLARIN. In order to achieve this the national CLaDA-BG coordinator, prof. Kiril Simov of the Institute of information and communication technologies at the Bulgarian Academy of Sciences and prof. Petya Osenova of Sofia University were regularly invited as guest lecturers. A staple of the programme at each of these events was the discussion on the problems featured in the lectures.

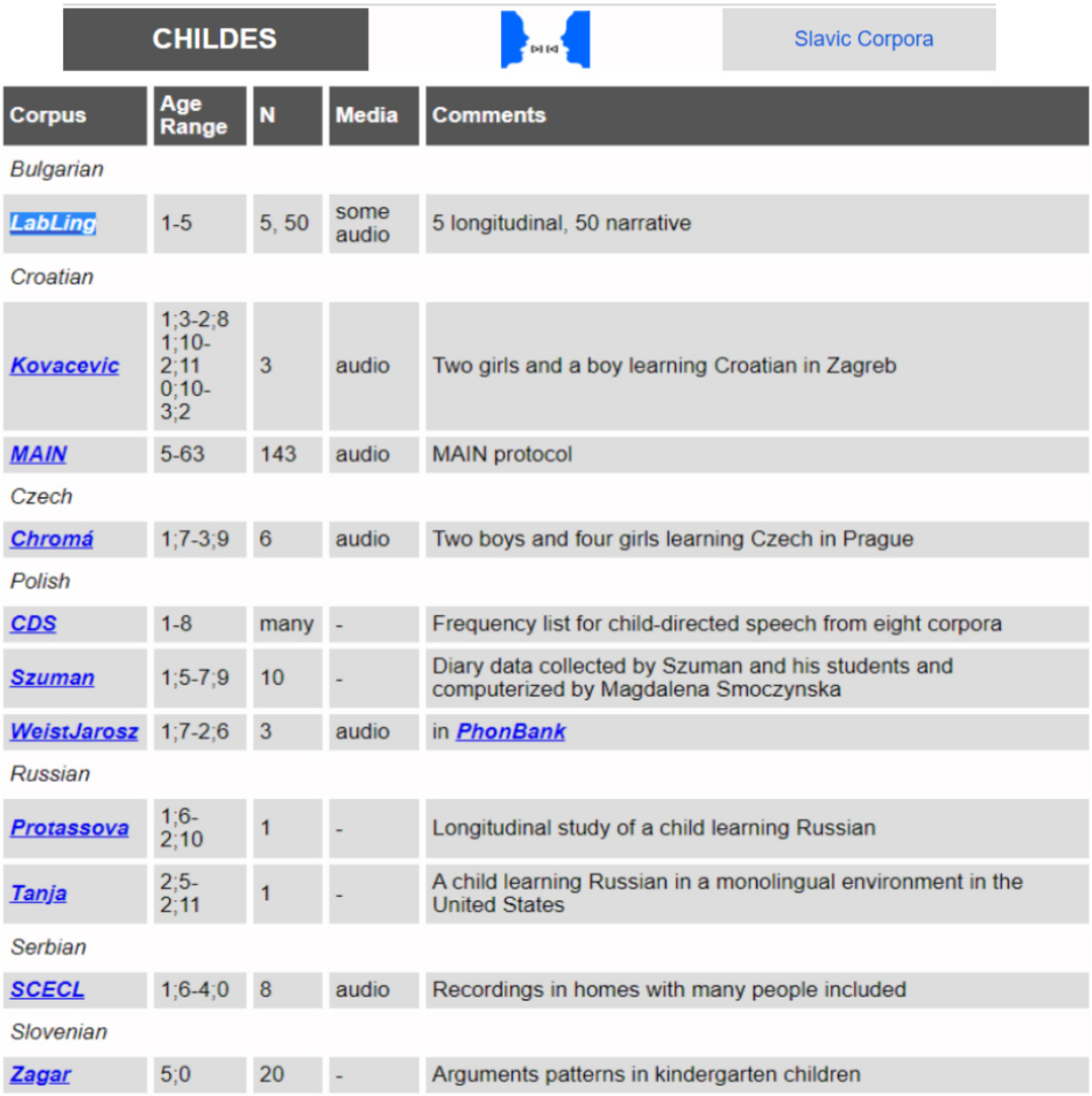
With the development of the CLaDA-BG infrastructure and the establishing of LABLING as a technological partner came inner-circle, more specialized seminars in a narrower format. They were related to the preparation of the participants in the project for work on various research activities as well as for working with various platforms. The knowledge and skills acquired in the process further broadened the professional skills of the students and helped their personal and language development. In this sense, this extracurricular activity can definitely be interpreted as a specific element of the university education process which has transitioned out of the classroom and into the research lab.

The specialized LABLING seminars held so far are related to the two priority research areas for the LABLNG team, namely the creation of collections of associative and child speech databases. For this purpose the undergraduate and PhD students were introduced to the resources and instruments of the respective platforms (LABLASS and CHILDES) and the necessary skills for working in this specific environment developed on this basis. In the context of this training of sorts the young people were divided into small groups to ensure excellent work feedback and the ability to independently fulfill research tasks. Along with the specific skills necessary for every linguist such as the ability to collect and systematize data, transcribe speech data in specific formats, annotate and code corpora etc., the young people were able to gather enough experience with working as a team, as well as to develop in themselves the associated personal qualities.

In conclusion it can be stated that before CLARIN entered the academic classroom with its resources, instruments and services the classroom itself had become a workshop of sorts for CLARIN where in the process of creating child speech corpora and associative collections the trainees acquired research competences and skills, as well as the self-esteem that their future products will return to them and their colleagues in the university auditorium.

2.1 The LABLASS Web-Based System and the Bulgarian LabLing Corpus in Teaching Linguistics at University

The pilot versions of the LABLASS web-based system and the Bulgarian LabLing Corpus are already a fact after two-years of work (2019-2020). They have been immediately included in the curricula of linguistics disciplines and the newly published textbook by one of the authors of the present article (Popova, 2020). Within the context of the importance of these two CLaDA-BG products, the following section will attempt to present the specific projections of their useful application in teaching Linguistics in some of the philological and pedagogical majors at Shumen University. The fall of 2020 saw the addition of the first Bulgarian child speech corpus (Bulgarian LabLing Corpus) to the Slavic languages section of the CHILDES database platform which includes data about the acquisition of numerous languages from various language families (see Fig. 1).

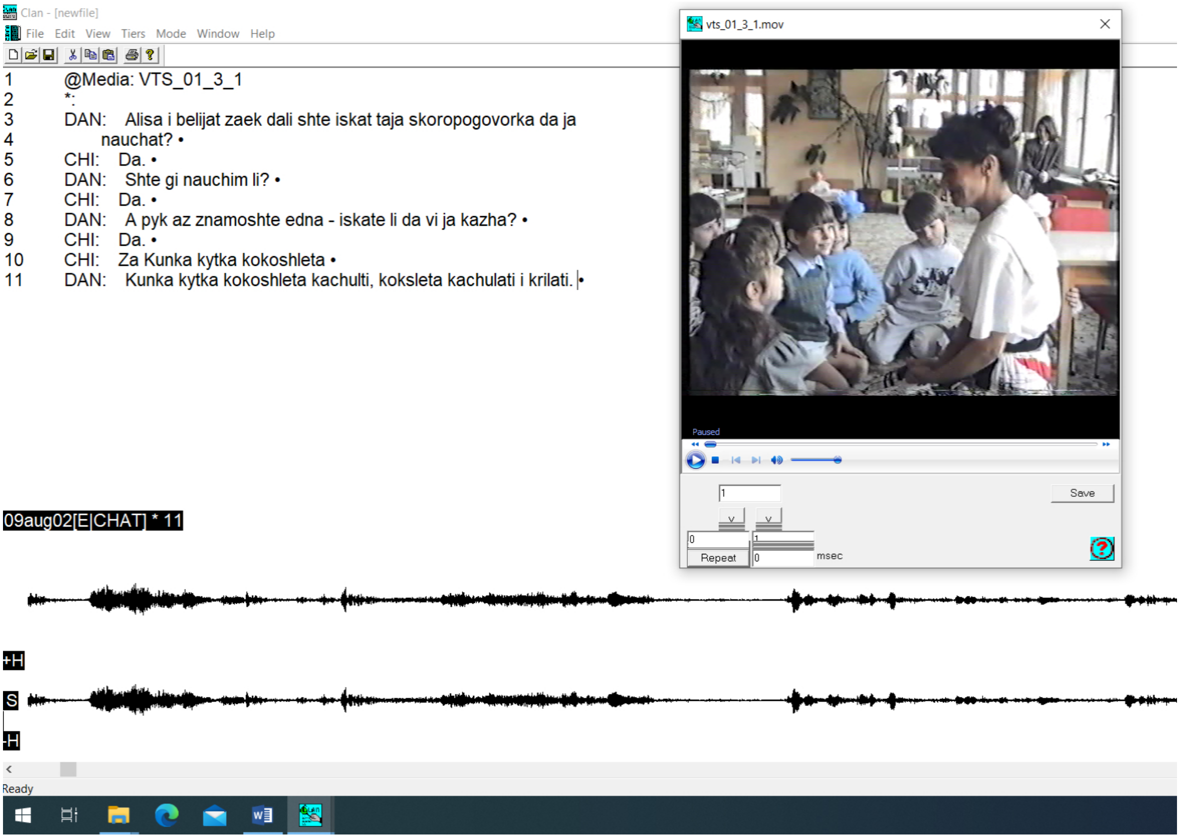


Corpus	Age Range	N	Media	Comments
Bulgarian				
LabLing	1-5	5, 50	some audio	5 longitudinal, 50 narrative
Croatian				
Kovacevic	1;3-2;8 1;10-2;11 0;10-3;2	3	audio	Two girls and a boy learning Croatian in Zagreb
MAIN	5-63	143	audio	MAIN protocol
Czech				
Chromá	1;7-3;9	6	audio	Two boys and four girls learning Czech in Prague
Polish				
CDS	1-8	many	-	Frequency list for child-directed speech from eight corpora
Szuman	1;5-7;9	10	-	Diary data collected by Szuman and his students and computerized by Magdalena Smoczynska
WeistJarosz	1;7-2;6	3	audio	in PhonBank
Russian				
Protassova	1;6-2;10	1	-	Longitudinal study of a child learning Russian
Tanja	2;5-2;11	1	-	A child learning Russian in a monolingual environment in the United States
Serbian				
SCECL	1;6-4;0	8	audio	Recordings in homes with many people included
Slovenian				
Zagar	5;0	20	-	Arguments patterns in kindergarten children

Figure 1. Bulgarian LabLing Corpus in the CHILDES Slavic Collection.

The published first corpus of children‘s speech (Bulgarian LabLing Corpus) is freely accessible for researchers at `https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html`. Its broad applicability is due to the fact that each of the transcripts includes data for identification of the research subjects (demographic and language parameters) and regarding the type of the corpus itself (longitudinal or cross-sectional). Meanwhile, with its addition to the common database of CHILDES, the abilities of the system for crosslinguistic research is enriched with another Slavic language. Additionally, the Bulgarian linguistic tradition is enriched with another universal applicable standard for researching language ontogenesis thanks to which researchers can make fast, exact and reliable juxtapositions with a large number of languages and build solid typologies and modern theories.

We also need to point out the unquestionable benefits of broadening the students‘ corpus competence by including knowledge about CHILDES as part of TalkBank – one of the most successful contemporary platforms for studying human speech behaviour. Thus the curricula of the disciplines Psycholinguistics, Foundations of language acquisition, Linguistics, Child Linguistics have been thematically extended which in turn improved the standard and the quality of independent student research in the form of course assignments and theses. In addition, the data from the Bulgarian LabLing Corpus are often used in the teaching process as their multimodality makes them useful and applicable in different demonstrations (see Fig. 2).



```
@Media: VTS_01_3_1
*:
DAN:   Alisa i belijat zaek dali shte iskat taja skoropogovorka da ja
         nauchat? •
CHI:   Da. •
DAN:   Shte gi nauchim li? •
CHI:   Da. •
DAN:   A pyk az znamoshte edna - iskate li da vi ja kazha? •
CHI:   Da. •
CHI:   Za Kunka kytka kokoshleta •
DAN:   Kunka kytka kokoshleta kachulti, koksleta kachulati i krilati. •
```

Figure 2. Multimodal data presentation.

The LABLASS web-based system developed within the CLaDA-BG project does not at this point provide free access to users although this is projected to happen by the end of 2021. Its usefulness for researchers is not limited to the availability of the associative collections of included lexicographic sources but is instead much broader and concerns providing the option for the user to create their own new dictionaries and to visualize and compare data from various sources. This web-based system has its place in the practical units of the disciplines dealing with topics such as the mental lexicon and language ontogenesis since word associations are an extremely important source of information in that respect.

In their classes, students can test their working hypotheses by comparing and analyzing published data, creating their own dictionaries. On this basis they develop their own research within the specific course projects. A very important positive result in this respect is the successful defense in 2019 of an MA thesis entitled “Specificities of the Vocabulary of the Bulgarian Native Speaker Nowadays. A Psycholinguistic Study”.

The success of independent student research is supported not only by the resources and instruments of the web-based system LABLASS and the CHILDES platform, but also by the additional unit in the curriculum presented in the textbook, *Psycholinguistics as Experimental Linguistics* (Popova, 2020). It is a workshop of sorts aimed at developing students’ skills for planning and conducting associative experiments and for gathering empirical data and structuring it as a corpus. Also undeniable is the importance of the updated curriculum content of the practical units as well as the academic guidance on the part of the professor and the inclusion of interdisciplinary scenarios for classes with the participation of specialists from different academic fields (a linguist and an IT specialist).

In order to illustrate the ideas mentioned above we can use as an example a pilot model for curricular work in Psycholinguistics with Special Pedagogy students at Shumen University during the winter semester of the 2020–2021 academic year. We shall observe the scenario which includes a cycle of classes supplying the students with knowledge and skills for planning and completing an independent study.

The first class was entitled “Theory and practice of the psycholinguistic experiment”. It was held on the 3rd of November, 2020. The interdisciplinary scenario with the participation of specialists from different academic fields – the psycholinguistics leader and the IT specialist guest lecturer (in fact one of the creators of the LABLASS web-based system) – first was successfully tested (see Fig. 3).

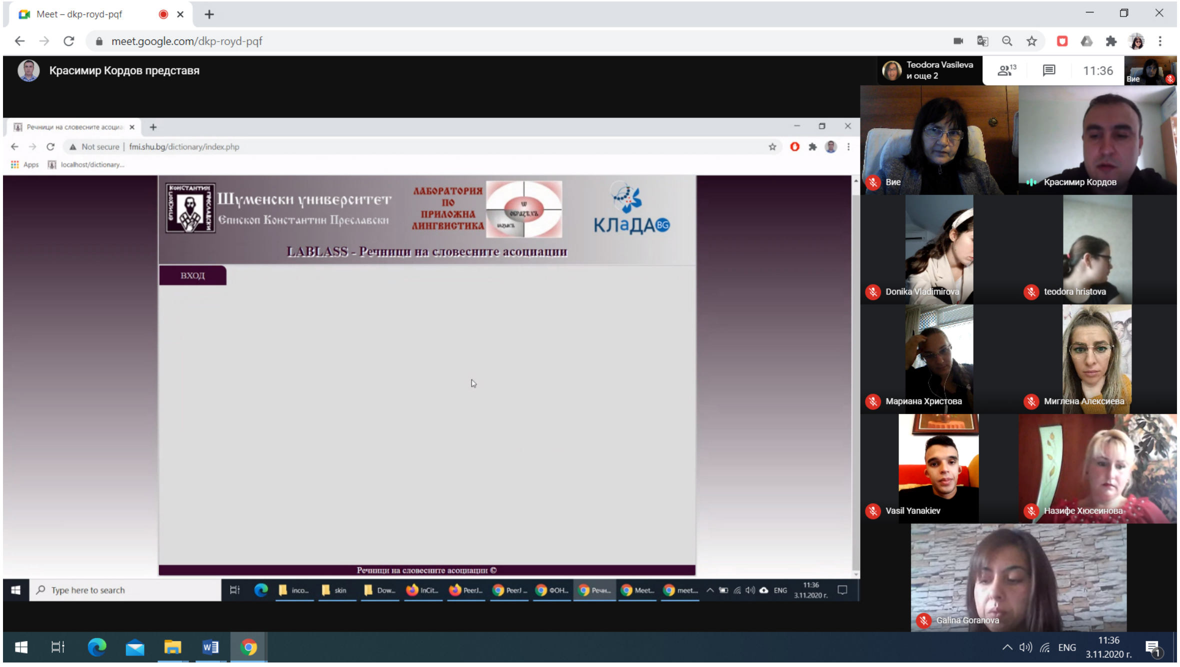


Figure 3. An interdisciplinary scenario class - 3.11.2020, Shumen University.

At the beginning of the class the professor with the leading role in the class, the psycholinguist, introduced the topic in accordance with the compulsory curriculum. After that the guest-lecturer (as illustrated above) was given the opportunity to present the functionalities of the newly-developed web-based system to the students. The discussion following the demonstration of the platform included the contributions of the student-participants in the CLaDA-BG project as well.

A similar scenario (albeit without a guest-lecturer) was employed during the next class dedicated to the importance of corpora and Corpus Linguistics for the professional capacity of the special pedagogue. The TALKBANK and CHILDES platforms were used as general illustration after which, within the context of introducing their functionalities and resources, the functionalities of the system were demonstrated.

After the first two classes came a discussion of the research intentions of each of the trainees with a view to the planning and implementing of an independent study.

This pilot cycle integrated within the Psycholinguistics course is aimed mainly at research techniques and work methods since they are the ones which create the opportunity for achieving the optimal balance between theoretical knowledge and practical skills in the process of education, as well as for the development and broadening of the competences of the trainees. In addition, the updating of the respective curriculum content with regard to the introduction to the pilot versions of LABLASS and the Bulgarian LabLing corpus enables students to be more active in the process, i.e. they are not handed down knowledge but instead co-discover it along with the teacher as academic advisor.

Similar was the scenario for the classes in Child Linguistics and Foundations of language acquisition for students of philology where the practical unit also stood out. In the Introduction to general linguistics course, however, the work on improving the corpus competences of the students was done only in the form of an addition of sorts to the curriculum content, and done only for some topics. Moreover, during the course of the specific classes, while working on particular linguistic case studies and in the completion of individual and group tasks for learning and research, the students were given the opportunity to test some of the functionalities of the web-based system LABLASS as well as those of the CHILDES platform.

3 Conclusion

The article attempted to demonstrate some of the positive effects of the *CLARIN in the Classroom* initiative, of the example of the pilot testing of the web-based system LABLASS and the first Bulgarian child speech corpus (Bulgarian LabLing Corpus) developed within the CLaDA-BG project for teaching linguistics to students of philology and special pedagogy at Shumen University. To summarize, the specific applications of these products in the university teaching practice concern mainly the updating of the curriculum content and the practical units of the specific linguistic disciplines, the creation of teaching resources, the introduction of interdisciplinary scenarios for classes with the participation of specialists from different academic fields, as well as the transitioning of the teaching process out of the class-room and into the research lab. The preliminary results include the improved theoretical and practical competences, language development, research curiosity and professional self-esteem directly reflected in the independent articles, presentations, thesis projects of students where the young people's interest in other CLARIN resources, instruments and services is already observable.

Acknowledgements

This research was partially funded by the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG, Grant number DO01-272/16.12.2019.

References

Simov K., Osenova P. 2020. Integrated Language and Knowledge Resources for CLaDA-BG. *Selected papers from the CLARIN Annual Conference 2019.* Linköping Electronic Conference Proceedings: 137–144.

Popova V. 2018. *Izledovatelskiyat podhod v obuchenieto po balgarski ezik mezhdu tradicionnoto i inovativnoto* - Bulgarian Language and Literature. 62(3) (in Bulgarian).

Popova V. 2020. Psiholingvistikata kato experimentalno ezikoznanie. *Shumen: Univ. izdatelstvo "Episkop Konstantin Preslavsky"* (in Bulgarian).