Adaptation of Swedish Transcription System for Spoken Language Analysis for Bulgarian

Krasimira Petrova[1], Krasimira Aleksova[2]

Sofia University "St.Kliment Ochridski", Department of Slavonic Languages,
[1]Chair of Russian Language, room 128; [2] Chair of Bulgarian Language
15, Tzar Osvoboditel Blvd. 1504 SOFIA, BULGARIA
tel. 003592 9308 206 or 93088 300; fax 003592 946 02 55 or 88 14 42
E- mail: krasi@slav.uni-sofia.bg; aleksova@slav.uni-sofia.bg

1. A bilateral project "Multimedia and Multimodal Spoken Language Corpora Analysis. Stage 1" with the partnership between the University of Götheborg (GU), Insitute for Linguistics, Sweden and Sofia University "St. Kliment Ohridski" (SU), Faculty of Slavic Studies, Bulgaria took place at Sofia University in spring semester in May 2000. Swedish contractors Biljana Martinovsky and Leif Grönquist (we thank them cordially!) introduced a course to the students, PhD students and teaching staff, and in collaboration with the Bulgarian partners worked out, fulfilled the objectives of the Stage 1 of the project (the description, report and samples of collected data can be viewed at http://www.ling.gu.se/~leifg/sofia/).

2. The aims and goals of this project are closely related to the main trends of research and teaching activities of lecturers and teachers of the Departments of Bulgarian language, Russian language and Methods of teaching at SU (see http://www.slav.uni-sofia.bg). The common problems of the collection, organisation, transcription and annotation of a corpus of Spoken Bulgarian from the different kinds of social activities with the help of modern multimedia technology were discussed. This seminar was one of the first steps to help us to start developing the Standardised corpus of Bulgarian Spoken language which can be used for future psycho- and sociolinguistics, inter- and cross-lingual research projects.

2.1. Taking part in this project benefits our work in the following directions:

2.1.1. Collection, organisation, transcription and annotation of a corpus of Spoken Bulgarian from different kinds of social activities (such as classrooms, business meetings, courtrooms) according to standards developed at GU.

2.1.2. Development of a course on corpus linguistics and discourse analysis for Bulgarian lecturers and graduating students.

2.1.3. Introducing and instructing the Bulgarian teachers and graduating students to the use of computational multimedia tools developed at GU for research and teaching purposes.

2.1.4. Publication of a manual for the use of the corpus on paper and on internet.

2.1.5. Development of a Standardized Corpus of Spoken Bulgarian (SCSB) language representing different areas of social life, which can be used for future sociolinguistic and inter-lingual research projects.

2.1.6. Updating the linguistic competence of Bulgarian teachers and researchers according to the conditions of the modern multimedia technology.

2.2. Despite the fact that colloquial Bulgarian has been broadly studied in traditional sociolinguistic manner there is no properly prepared, transcribed and coded corpus of

spoken Bulgarian at all. Individual researchers have gathered samples of data on audio tapes but these data are not organized, described or transcribed according to international standards (see, for example Cvetanka Nikolova's corpus (50k tokens): http://www.hf.uio.no/east/bulg/mat/Nikolova/; Krasimira Alexova's corpus (100k tokens) http://www.hf.uio.no/east/bulg/mat/Aleksova/; transcripts made by Ivanka Mavrodieva from recordings made at the Sociolinguistics Laboratory at Sofia University of broadcasts from the debates of the 7th Great National Assembly on 31 October, 1990 at http://www.hf.uio.no/east/bulg/mat/Parliament/;) . The accomplishment of this task enables the use of Bulgarian in cross-linguistic comparative work and is to be added to the multilingual corpora of various spoken languages collected at the Department of Linguistics, GU. It gives an access to Bulgarian linguists to knowledge in contemporary linguistic development in the areas of pragmatics, discourse analysis, conversational analysis, as well as computational linguistics, which facilitates more active research base on Bulgarian and intercultural linguistic aspects of communication. Both Bulgarian and Swedish are carried by about 9 million speakers and are statistically evaluated as dying languages. Their historical development has many similar trends despite the difference in linguistic grouping. In difference from other Slavic languages, such as Russian and Polish, Bulgarian spoken language interaction has not been studied empirically enough.

3. The project is to be divided into two stages:
3.1.Stage 1:
3.1.1. preparation, organization and distribution of tasks
3.1.2. collection of data
3.1.3. course in multimedia and multimodal communication studies
3.1.4. transcription and coding

3.2. Stage 2: analysis of interactive features in different social activities: courtroom, classroom and business interaction, communication on political matters

3.3. Fulfilled tasks on the Stage 1 are the following (see http://www.ling.gu.se/~leifg/sofia/doc/Sofia-report.html):

1. Created course compendium
2. Lectures on theory of interaction analysis
3. Recorded and digitalized 4 video films and 5 audio tapes
4. Installation of LINUX on 4 computers
5. Adaptation of TRASA and related instruments to LINUX
6. Installation of TRASA on 4 computers
7. Formulation of a Bulgarian manual for TRASA
8. Instructions and exercises on UNIX
9. Instructions and exercises on TRASA
10. Formulation of transcription standard for Bulgarian (TSB)
11. Formulation of modified orthography standard for Bulgarian (MOSB)
12. Transcription of 5 recordings
13. Statistical analysis of 5 transcriptions
14. Conversation analysis session
15. Installation of MULTITOOL on one computer and demonstration
16. Formulation of 4 essays

4. Here we focus on the modification of orthography standards for Bulgarian.
4.1. The both forms of the language – spoken and written – have their specific features but both of them serve the process of communication. We can observe dramatical diversion between these two forms, and this fact causes some difficulties and conventionalities in the process of "converting" the spoken form into written. Within spoken form we can differentiate between full standard orthoepic norms and conversational or dialect pronunciation. Similarly, two forms of written language exist – transcription and orthography. Transcription precisely reflects all the peculiarities of spoken language. Orthography is a fixation of the speech but there is no complete identity between phonetic and graphemic systems of a certain language. The relation and correspondence between phonemic and graphical systems is sustained by dynamic but not static balance, as the phonemic system is dynamic and being constantly developed, whereas the (ortho)graphic system is tend to be constant, static and conservative [see Tilkov et all, 1982:285]. Three basic principles underlie Bulgarian orthography: phonetic, morphologic (or etymological) and historical ones [see also there:291].
4.2. The adopted and modified Swedish standards for transcription of the spoken language texts (for multimedia tools TRASA and TRACTOR – see Nivre (1998) combine both types of putting spoken language into written – transcription and orthography. Transcription literally reflects the speech of communicants; orthography presents the words in standard written form in order to identify them correctly in case of diversion from the standard orthoepic norm, homophony, homonymy, etc. We use curly brackets {}, and put what is pronounced in the first place in them and after the colon we put the orthographic norm in order to keep the identity of the morpheme. There are several types of inconformity, non-correspondence between the pronunciation and orthography of the wordforms.

4.2.1. devocalization of the voiced consonants at the end of the word and in front of the unvoiced consonants in the middle of the word. Example:
http://www.ling.gu.se/~leifg/sofia/doc/zvetana/recact2.html
$A: И а**{с:з}** ставам пре{д} стави си1 / отивам / тихо / отварям вратата / Веса въ**{ф:в}** това време се съблича / нош{т}ницата си облича и като се стресн{А:а} < тая пуста жена > …
$D: Тя ш{т}е си2 и**{с:з}** кара акъла ма
$B: и тя ф{в}се по**{т:д}**скача Веса / а ти к{ак}во си каз{А:а}ла на майка си2
$C:тя съ**{ф:в}**сем така

4.2.2. vocalization of the unvoiced consonants in the middle of the word (the orphoepic norm is like in 4.2.1. above) – Example: вра**{б:п}**че

4.2.3. vocalization of the unvoiced consonants in front of voiced consonant
http://www.ling.gu.se/~leifg/sofia/doc/Ivelina%26Nevena/ive.html
$A: [6Получавъш]6 двата о**{д:т}**говора и разбираш кой е счупенийъ стол
$B: Не1, получаваш един о**{д:т}**говор,

4.2.4. ellizion of a phomene / phonemes is marked by zero after the colon in brackets (*compare with 7) – it should be {0:__ }) or put simply in brackets:

$B:[14 разбираш]14 ли / штото при / пружинката / и тя ф{в}се по{т:д}скача Веса / а ти к**{ак:0}**во си каз{А:а}ла на майка си2
$B: Единийъ електрически стол е счупен, другийъ е здра{ф:в}// и трьа**{бв}**а да откри{й:е}ш
$A: Добре/ и те съ пре{т:д} столовете и **{з}**начи трьабва да им зададеш един въпрос

4.2.5. replacement of a phoneme/ some phonemes by the other/ others:
$A:а1 стига ка**{й:зва}** превземки / какво беше
$C:тя съ{ф:в}сем така

4.2.6. qualitative reduction of the vocals – a dialect feature vs. quantitative reduction as a orphoepic norm (underlined in the example below):
$A: Помниш ли **{у:о}**на**йъ** задача за:/ к{акв:0}о беш**{и:е}**/, з**{ъъ:а}**// за двамат**ъ** **{у:о}**съдени на смърт на електрически**йъ** стол. Как беш**{и:е}** тьа?/

4.2.7. insertion, parenthesis is marked by zero after the colon in brackets:
пер**{е:0}**спектива

4.2.8. letter **щ** is represented by **шт**
$B: Не2 съ испържени о**шт**е.>1
$A:да1 бе / за**шт**о [14 така]14

4.2.9. letter **я** is represented by **ьа, ьъ, йъ, йа:**
$A: Стиг{ъ:а} де. Как беш{и:е} т**ьа**?/
$B: Не2 **{е:йа}** помн**ьъ**. Припомни ми **йъ**.

4.2.10. letter **ю** is represented by **йу** or **ьу:** ап{б}сол**ьу**тно

4.2.11. soft consonant is marked by **ь** after it combined with replacement:
$A: [9 Значи ти ако... Аа]9, чакай малко с**{ь:ег}**а ти питаш само едини**йъ** човек

4.2.12. any replacement: с**{п:па}**ъ, see also example in 4.2.11. above

4.2.13. incorrect replacement of the accent is marked by a capital letter:
http://www.ling.gu.se/~leifg/sofia/doc/zvetana/recact2.html
$A: к**а**то ги ококор**{И:и}** ония очи и падн**{А:а}** на креват**а** /

4.2.14. different degrees of reduction of the vowels according to the orthoepic norms of Bulgarian are not represented in the transcriptions (vowels are underlined in example 4.2.13.); any irregularity of these rules are represented (see example 4.2.12.)

4.2.15. as in the phonetic transcription, inflection for verbs, present tense, 1-st person singular and 3-rd person plural, and also full and short form of the definite article for nouns, adjectives and participles are written by **ъ.**
$A:викам колко е час**{ъ:а}** бе/

$A: Чакай малко лъжец**ъ** ако стои пред счупени**йъ** стол ш{те} ти кажи не1 не2 сто**йъ** пред счупени**йъ** стол

5. The concrete results were twofold: a) the establishment of the beginnings of a Standardized Corpus of Spoken Bulgarian; b) the increase of Bulgarian students' and lecturer's competence in the area of multimedia and its application in corpus based language studies. The first stage of this project contributes to the raising the awareness and competence in this field and enthusiasm for further building up (see http://www.larflast.bas.bg/balric/index.html), to encourage creating a consortium of researchers and teachers, unifying their efforts and working as a team across different institutions – academy of science, university, etc (see http://www.lml.bas.bg, http://www.BulTreeBank.org), also the recently established Master's Program for Computational Humanitarian Studies at SU (see http://www.slav.uni-sofia.bg/Pages/COMHU.htm) which goal is to provide graduates theoretical and practical training in the field of Computational Linguistics and Natural Language Processing.

REFERENCES

Nivre J. (1998). *Transcription Standard.* Semantics and Spoken Language, Department of Linguistics, Göteborg University, Version 6 – Gotheburgh papers in theoretical lingusitics. Göteborg.

Tilkov D., ed. (1982). *Gramatika na savremennija bulgarski knizhoven ezik. I. Fonetika. (= Grammar of contemporary standard Bulgarian language. Volume 1. Phonetics).* Sofia, Bulgarian Academy of Science Publishing house.