Computer Applications in Slavic Studies

Proceedings of *Azbuky.Net*

International Conference and Workshop

24–27 October 2005, Sofia, Bulgaria

Computer Processing of Modern Languages

Colloquial Bulgarian on the Web

Yovka Tisheva, Marina Dzhonova

(Faculty of Slavic Studies, Department of Bulgarian Language

"St. Kliment Ohridski" University, Sofia, Bulgaria)

Abstract: In this paper we discuss the description, analysis and possibilities for an edition with computer tools of data collected from Bulgarian colloquial speech. The colloquial speech being the most dynamic part of Contemporary Bulgarian needs specific means for an adequate description and analyses. Special attention should be paid to its particular features: ellipsis, contractions of forms and phrases, overlapping, multiplication of constituents and phrases. Moreover, since the spoken language is mostly performed as a dialogue we claim that analyses at the level of discourse and communication must be implemented as well. This set of characteristics determines the specificity of colloquial (spoken) data and imposes concrete decisions concerning their description, edition and annotation. There are already collections of Bulgarian colloquial data published on the Internet (www.hf.uio.no/east/bulg/mat, www.slav.uni-sofia.bg/~bgspeech/indexbg.htm). However, they are not processed in the way that could allow their treatment as annotated corpora. In this paper we outline the most salient syntactic features of spoken (colloquial) Bulgarian in order to support and benefit the future work on coding/processing language data at different levels. Our proposals could also have an important implication in the future work at the level of spoken dialogue and discourse, computational and corpora linguistics or speech processing.

I. Introduction

Much work has been done recently on creating Bulgarian language corpora of various types. All of them, however, represent one of the forms of contempo-

rary Bulgarian associated with standard language and written texts. In order to make a corpus which is more representative for a real language communication and "open" for different tasks and users it is worth trying to integrate the knowledge about the standard language system and its applications in the speech acts. Data of spoken language could be published on the web as correlates to the written texts (transcriptions/text files) or as audio/video files. In this paper we will discuss only the characteristics of data published as transcriptions, because this is the main format used for representation of spoken Bulgarian.

So far the data of spoken Bulgarian available on the web have been used mainly for "pure" linguistic purposes – to find examples or judgments for linguistic analyses. Well-ordered information about the main features of spoken Bulgarian which could be accessible even for users without preliminary knowledge of Bulgarian will encourage more scholars to include Bulgarian data in cross-linguistic comparative works. Since each corpus generally consists of two parts: a collection of texts (of different sorts, selected and visualized in different ways) and a header (the meta-information about the textual part) the information for the latter could enrich significantly the linguistic information and increase the interest in the colloquial data. A multimodal corpus of transcripts, audio and video files could find appliance in theories of dialogue and communication or as training data for language teaching, spoken dialogue systems, speech synthesis, etc.

Once the data of spoken language are collected and stored as transcripts and/or digitized archives an important question arises: which features of spoken Bulgarian might be annotated in such a data base? Linguistic analyses might give some solid reasons for motivating the selection of features. If one wants to create a corpus representative for the main spoken language phenomena, profound knowledge about them will be the point to start annotation. In this paper we consider that the theoretical analyses of contemporary spoken Bulgarian provide us with such list of representative features. On the other hand, tagging depends on the research agenda. Thus we have to account for the interaction between the annotation tasks and different researcher's aims. We would like to build the corpus as reach as possible, but we have to start with most basic data organized in the way that everyone could process them. The aims of annotation cover linguistic purposes in the first place, but they also give the possibility for the future implementation of formulating theories of dialogue, building models of communication, social interaction, etc.

The main goal of this paper is to outline the most salient phonological, grammatical and (some) discourse features of spoken (colloquial) Bulgarian in order to support and benefit the future work on annotating, coding and processing spoken language data. In the rest of the paper we first represent the resources of spoken Bulgarian available on the web. Then we show a model for description and analysis of linguistic phenomena of spoken Bulgarian (within the framework of linguistics). We also discuss the specificity of levels of annotation in order to build a corpus of spoken Bulgarian. Finally, we give some proposals for future edition with computer tools of spoken Bulgarian data at both audio- (digitized data base) and transcriptional levels.

II. Resources available on the web

There are few data bases of spoken Bulgarian published and accessible on the web. We will present them chronologically, i.e. in order of their publishing. Thanks to the transcripts published by Kjetil Raa Hauge on the University of Oslo web site (www.hf.uio.no/east/bulg/mat) – Spoken Bulgarian collected by Cv. Nikolova; Spoken Bulgarian collected by Kr. Aleksova; Bulgarian Parliament debates collected by I. Mavrodieva, and ICQ conversations collected by M. Dzhonova, a new source of information about Bulgarian appeared.
The data collected by Cv. Nikolova were recorded in randomly selected places (shops, streetcars, offices, homes) during the years 1975 to 1977. The text files (normalized transcriptions within the standard orthography) contain only the sentences uttered by the informants, without indication of speakers' identities and turn changes. They are therefore best suited for investigations of phenomena at the level of syntax, not dialogues or social/communicative interactions. Aleksova's corpus provides better material for investigations of discourse phenomena. The data collected by Kr. Aleksova consist of transcribed conversations made in family contexts during the years 1989–1990. The data collected by I. Mavrodieva contain transcripts of broadcasts from the debates of the 7th Great National Assembly on 31 October 1990. The latest database contains conversations under the IRC (Internet Relay Protocol) in Bulgarian chatrooms recorded in 2001.

The transcripts published by K. Hauge represent the language situation in Bulgaria during the last century. New data are needed in order to show the speech of new generation, mainly of the young people, and also to cover new language situations, not only dialogues in the family. "No corpus can be fully representative sample of the whole language, but such collection can at least be designed to represent major dimensions of language variations."(Stubbs

2001: 305). The transcripts available are considered insufficient for building a corpus of spoken Bulgarian, because they do not contain any information about the lexical or grammatical forms of the words, which do not correspond to the standard rules. The phenomena typical for spoken Bulgarian are well represented, but once being taken out of the transcripts, forms and phrases need normalization. The latter, in fact, is exactly what most of the users do in order to implement these data in their own research.

The first attempt to build digital corpora of spoken Bulgarian was made in 2000–2001. Due to a joint project between Sofia University and Gothenburg University "Multimedia and Research on Multimodal Social Interaction" (www.ling.gu.se/~leifg/sofia) a small database containing audio and video recordings and their transcripts was created. The efforts of the project team did not attain the desired result, because the annotation used for these transcripts was applicable only for specific computer tools – TRASA and TRACTOR.

A team from the Faculty of Slavic Studies at the Sofia University started working on a web site for Spoken Bulgarian (named Bgspeech). At this stage, it contains both transcripts of spoken data and publications on spoken Bulgarian. The team has also collected audio and video recordings of dialogs representative for different types of social interaction. Some of them have also transcripts.

To summarize this short review, the publications of spoken Bulgarian data on the Internet could be defined as data collections or data bases. Our aim is to create a corpus of spoken Bulgarian, which will be not only a large collection of computer-readable texts and audio/video recordings of different types of speech acts. This paper addresses questions such as which are the relevant factors when making and using transcripts, which formats and encoding standards are most appropriate for spoken data, etc. It is considered that "TEI (or something like it) is really the only workable basis for standardization. A standard of this type seeks to provide a mechanism (via markup conventions) for systematic encoding of data, such that the data can be flexibly reformatted later in various ways..." (Edwards 2001: 343). As for "content-based" decisions which cannot be adjusted by computer tools at this stage we propose that the corpus might contain two types of information: general linguistic information (concerning the lexical form of the items and their correlation with the standard lexicon; the POS tagging) and information about particular features (ellipsis, contractions of forms and phrases, overlapping, repetition of

constituents and phrases, etc.). The latter has features that are generally relevant for spoken languages and are well presented in the data from modern spoken Bulgarian. We propose a separate level of annotation for pragmatic and discourse markers in order to account for the multifunctionality of these linguistic items. The pragmatic and discourse markers are included in the lexicon (some of them will even get their POS tags), although they operate mainly at the level of communication and dialogue interaction.

III. Spoken Bulgarian – main features

Since our aim is to design (a model for) a corpus representing the specificity of spoken variation of Bulgarian we start with an introduction of features relevant for annotation schema. Some of the domains in which data collections differ most often are related exactly with levels and units of analyses and encoding. The main problem in encoding lexical items is whether to preserve the nuances of pronunciation, or to use the standard orthography. Here this will not be discussed, but some specific features on the level of phonology and morphology will be presented with the implication that these features should be encoded in the transcripts.

1. Phonetics

At the level of phonology the most distinct features of spoken language are ellipsis and prosodic patterns, which differ from the standard pronunciation. These phenomena are widely studied within the framework of Bulgarian phonetics and dialectology. What is important for the annotation of spoken Bulgarian data is the fact that some phonetic changes have influence on morphological level, particularly the unification of verb inflections and definite articles. These two processes are typical for contemporary spoken Bulgarian. In example (1) the verb from II conjugation type (i-group) gets the inflexion for 1p.pl. present *–me*, which is in fact for the III conjugation type:

(1) ne može da go prevrâštame v učeničeski turizâm samo zaštoto trjabva da pestime
We can't transform this into school tourism because we have to save money

In example (2) a very common generalization of the inflexion *-âx* for 1p.sg. aorist (used instead of standard inflection *-ox*) is illustrated:

(2) otidâx na start i kato otidâx na start se pojavi tva
I pressed the start button and when I did it this appeared.

The process of unification can be observed not only in the verbal paradigm. Example (3) represents the unification of formant *–â* for definite article not only for masculine gender, but also for feminine gender:

(3) do minus 6 sa padnali temperaturite prez noštâ
The temperature fell down to 6°C at night.

Even if the standard orthography is chosen for transcripts, these cases of unification should be encoded, because they show not a personal manner of speaking, but one of the most active processes in contemporary spoken Bulgarian.

2. Morphosyntax

In the previous paragraph we claimed that the distinctive phonological changes in spoken Bulgarian may have significant influence on the level of morphology. However, the changes traditionally defined as morphological, could be fully observed and explained only at the level of the sentences (utterances). In fact, the "border zones" between different levels are the less studied area of spoken Bulgarian, but the most distinctive one. Most of the features, representative for spoken Bulgarian, pertain precisely to this mixed group (e.g. doubling of elements, unification of syntactic tools, use of pronouns from different groups, drop of preposition *na*, etc.). And last, but not least there are some markers for spoken Bulgarian, which at first seems to be acting at the level of syntax, but the effect of their usage can be captured on the discourse level – overlaps, speech repairs, incomplete utterances, utterance boundaries.

The clitic doubling, which is found in all varieties of contemporary Bulgarian is one of the most studied syntactic phenomena. We argue that the doubling is the mechanism, which determines to a very large extend the structure of the syntactic units from spoken data. Two types of doubling could be defined:

– categorical – when one grammatical (functional) category is marked in the sentence,by multiple tools;
– structural – doubling of syntactic arguments.

Examples like (4) show typical case of doubling at the categorical level – double definiteness, expressed by the demonstrative *tija* and the definite article for plural *–te*:

(4) toj gi prati tija dvete momčeta tam
He sent there those two girls

For standard Bulgarian double definiteness is considered ungrammatical. In spoken Bulgarian, however, such constructions are very frequently observed.

The doubling of arguments could be characterized as a phenomenon pertaining to the structural level of analysis. There are two types of argument doubling in contemporary Bulgarian – doubling of the subject, shown in (5), and object doubling, shown in examples like (4). The construction with subject doubling consist of a noun or NP *katolicite* and nominative personal pronoun *te*. NP and the pronoun are coreferent:

(5) te nali katolicite taka pravjat
Catholics, they do it this way.

This construction is considered to be one of the mechanisms for topicalisation in Bulgarian. The personal pronoun is always for 3p. (sg. or pl.) and could be preposed with respect to the verb as it is in the case with *te* in (5) or to occupy a posverbal position like in (6):

(6) katolicite taka pravjat te

Usually the NP coreferential with the nominative pronoun is separated from the rest of the sentence by pauses of different length. Therefore the syntactic encoding of constructions with doubled subject should be also linked to phonology (marking the pauses).

The object doubling is known as *clitic reduplication* due to the fact that the doubling personal pronoun is a clitic. The phenomenon of the clitic reduplication of the object refers to the co-existence of a short (clitic) pronoun and of an associated non-clitic personal pronoun or of an object noun phrase. This discourse-syntactic mechanism is characteristic for all the varieties of Bulgarian: "apart from the literary language where it appears in some particular syntactic environments, it is extremely frequent in the colloquial style" (Krapova, Tisheva 2006: 415).

The clitic reduplication is not a unitary phenomenon neither in syntactic, nor in pragmatic terms, but rather, it appears as obligatory or optional, depending on factors such as the position of the object w.r.t. verb (pre- or postverbal), the type of clausal predicate, or the particular semantico-pragmatic function realized by the reduplicative mechanism. In (4) the accusative clitic *gi* is preposed with respect to the verb. It is coreferential with the direct object NP *tija dvete momčeta* and agrees with it in gender and number.

Although the doubling of arguments seems to be a syntactic mechanism, its real effect is actually performed at the pragmatic level, because it marks the specific part of information carried by the sentence (sentential Topic or Theme). We claim that even at the very early stage of annotation doubled structures could be marked. This will clarify the POS tagging since there will be two or more coreferential elements (which share also their agreement features) in the sentence. Different models of reduplication influence also the word order of the utterance.

As it was stated in the previous part, the unification takes place at the level of verbal and noun inflectional morphology. We consider the unification of syntactic tools to be also one of the most salient syntactic properties of spoken Bulgarian. Representative examples for this process come from the unified complementizer *deto* (a phenomenon observed in all Balkan languages). The morphology of standard Bulgarian classifies *deto* as a relative pronoun or an adverb. The following examples show that *deto* could be found in the wider contexts than relatives. In (6) *deto* substitutes the complementizer *če* 'that' and it could also mean the reason for subjects emotion and thus to substitute *zaštoto* 'because'.

(6) toj sega go e jad deto tja jade
Now he is mad that/because she is eating

In (7) *deto* appears in its contracted form *det* and stands for the locative adverb *kâdeto* 'where'.

(7) na taja spirka det slizame li?
At that bus stop where we get off?

The multifunctionality of *deto* also supports the very frequent use of this complementizer in colloquial Bulgarian. Apart from some adverb such as the locative *kâdeto* and the causative *zaštoto* the complementizer *deto* can easily replace relatives for persons/objects like *kojto, kojato, koeto, koito.* While the agreement in gender and number is needed between the relative pronouns and the coreferential NP, *deto* does not require any selection of grammar forms: it can stand for any antecedent. The representation of this lexeme in the annotation schema has to account for its specificity.

To a sertain extend the process of unification of morphosyntactic tools could be observed in colloquial examples with the missing preposition *na* (analysed in the literature as *na*-drop). Modern Bulgarian has two types of personal pro-

nouminal forms for accusative (direct object) and dative (indirect object): long tonic forms and short (clitic) forms. The tonic forms for both cases in fact coinside and the difference between direct and indirect objects is marked by using the preposition *na* (e.g. *mene, nego* for Acc. - *na mene, na nego* for Dat.). With some predicats (mainly predicats of state which require also the presence of a clitic to mark Acc. or Dat.) speakers tend to omit the preposition *na* as an overt marker for the syntactic function of the pronoun:

(8) nego pufi li mu vikat ?
Do they call him Pufy?

(9) men pâk mi j xubavo
But I feel good.

The process started from the constructions with personal pronouns, but same patterns with *na*-drop coudld be observed aslo with NPs. The omission of formal marker for indirect object may lead to some misinterpretations at the level of syntax and effects on the annotation of syntactic functions. Taking into consideration the ambiguities of forms, it would be appropriate to mark at a very early stage of annotation both morphological and syntactic features of pronouns and NPs found in *na*-drop constructions.

3. Syntax and discourse

At the level of syntax the question of the data structure and of the analyses units arises. Since the corpus is considered to be a *re*presetation of language data (here, from spoken Bulgarian), the main aim when building it is to preserve particular features of the language variation from which the data have been collected. Logically, for corpora of standard (written) Bulgarian the structure of written texts will be kept: chapters, paragraphs, lines will be the units to analyze. The dialogue (of at least two turn-takes or exchanges) or the polilogue are the usual forms of real language communication; monologs or long stories of one of the speakers are very seldom found in our data base. Obviously these structural units should be maintained in the corpus of spoken data (unlike the tradition in Bulgarian dialectology which is to arrange transcripts as unified texts for each speaker).

The dialogue linearizing is the process of segmentation of the speech of each speaker into smaller units with a single ordering of them. Within different theoretical frameworks or different purposes of analyses different subdivisions of texts are possible: episodes (speakers talk about one particular topic); turns (a word/phrase produced by one of the speakers without any influence

from other speakers), utterances (words/phrases corresponding structurally to a single sentence), prosodical/intonational units (a unit bounded by pauses and/or having a completed intonational contour), etc. The transcripts of spoken data give information about lexicon, grammar and some pragmatic aspects of this variation of Bulgarian. We consider these dimensions of the spoken texts to be well represented by choosing turns to be the first stage of the process of linearizing. Each turn consists of at least one utterance. In case of sequence of utterances produced by the same speaker (without interruptions from others) they will be placed within one turn.

There are different transcript formats for arranging the turns of the speakers:
– separate columns (column-based format) – gives a good vizualization of control and dominance during the conversation;
– single column (vertical format) – shows interdependence and equal dominance. (see also Edwards 2001: 326–329)

For our data we have adopted the vertical format because it implies the symmetry between speakers and gives an impression of how, from top to bottom, the discourse was built. The categories concerned with turn transition include "short pauses between one speaker and the next, interruption by the second speaker, and simultaneous talk (overlap)" (Edwards 2001: 328). Different parts of the utterances produced by the speakers could overlap. Most frequent is the case when the end (a phrase or a word at the end) of the first utterance overlaps with the start of the next utterance. As it could be seen from example (10), the end of the utterance of speaker A coincides with the start of the utterance of B. In the vertical format chosen here overlaps are marked by brackets:

(10) C: kvo be? čakaj
A: ni moa [az njaam]
B: [men] pâk mi j hubavu

More problematic to encode are the cases like those shown in (11). The words uttered by B form the "background" for the others. The whole utterance of C overlaps with the same phrase said by B. Shortly after C finished his turn, A started talking simultaneously with B. Then A stopped (the pause is signaled by // in the example) and started talking again after B ended his turn:

(11)
A: astra be/ astra
B: sin nadpis [astra] astra obikalja [ekipaža opa opa xvanaxa] momčeto sâs gramaža
C: [astra]
A: [ekipaža opa opa xvanaxa] // ne učastvaj v târg za koli

Examples like (11) show also that overlapping is one of the crucial problems for transcribers and annotators. The overlaps may differ significantly due to the time (if working with audio file) or place (in the transcript) in the utterance. The mid-utterance overlaps are less problematic to detect, but many questions arise when tagging them. In order to visualize them in the best way the layout of overlaps could be made different from the rest of the text: the column format could be replaced by a multilinear format, as we tried to show in (11). We do not reorder the overlapping speech, but suggest that the linearization of the transcribed dialogue could be shown better if the exact timing of overlaps is used to detect/annotate them.

The turns could be divided into "intonational phrases, which end with an acoustically signaled boundary tone. Intonational phrases have acoustic correlates such as the presence of silence."(Heeman and Allen 1997: 258). We though consider the utterances to be the parts, which the turns consist of. The utterance is a group formed by the consecutive words by a single speaker in which there is a minimal effect from the other speaker(s). One utterance may contain more than one sentence (e.g. complex sentences).

Marking utterance boundaries is one of the main problems at the level of syntactic units. It is relatively easy to detect the start of the utterance: it is either the start of the file or the end of utterance of the previous speaker (the cases of overlaps show more specific starts, which were already discussed). More problematic is to detect the end of the utterance mainly because of the cases of incompletion. The utterances could be incomplete either prosodically, syntactically or semantically. Our examples show mostly the cases of syntactic incompletion. As for the lacks at the level of semantics, in some cases, like in example (12), the utterance of the next speaker could be the "filler" for the missing part of the first utterance:

(12) A: i trimata koito bjahme na
A: The three of us who were at
B: na intervjuto
B: To the interveiw

By definition the spoken dialogues are spontaneous; the communicators sometimes start speaking before they are sure of what they want to say. Often the speakers abandon what they were saying; in the transcripts that part will be without any contribution to the rest of the dialogue (see e.g. (10), the utterance of A). The speakers might also need to go back and repair what was just said. Disfluencies where the speakers go back and change or repeat something they have just said are labeled as speech repairs. Whatever the reason for this is, the speech repairs are a normal occurrence in spoken dialogues. Interestingly, they tend to have a standard form. Thanks to that the hearers are able to effortlessly understand speech with repairs in it.

Heeman and Allen (1997) define 3 types of speech repairs:
– fresh starts – the speaker abandons the current utterance and starts again; the abandonment is acoustically signaled by pause/silence, fillers or interjections
– modification repairs – a editing phrase follows immediately the one which the speaker wants to repair
– abridged repairs – consist solely of an editing term which comes after a pause/silence or a filler

So far, in our data base we have paid more attention to the fresh starts (false starts in our annotation list):

(13) i meždu drugoto az (FS) â toj ima sa imen den
And by the way I (FS) hmm now he has a name day

The speech repairs could be annotated as based on preliminary analyses of linguistic features –they are usually marked by a short pause between two semantically connected phrases, a pragmatic or a discourse marker (in example (13) the multiword marker *meždu drugoto*) or fillers (in (13) schwa) before the speech repair.

In the spoken language pragmatic markers are one of the elements at the syntactic level of annotation, which cannot be ignored or regarded as noise. Although they are non-propositional sentence parts, they have to find appropriate encoding, because they signal the speaker's potential communicative intentions.

Fraser (1996) defines 4 types of pragmatic markers:
1. Basic markers – give information about speaker's basic communicative intentions. Performative expressions, illocutionary force markers, pragmatic idioms, tag questions, etc. function as basic markers. Interestingly, these phrases are more or less obligatory in the sentence structure.
2. Commentary pragmatic markers – optional phrases, which provide comments to the basic message.
3. Parallel markers – they consists of an entire message in addition to the basic one (vocative markers, solidarity markers, speaker's displeasure markers).
4. Discourse markers –mark the relationship of the basic message to the foregoing discourse. They provide instructions to addressee on how the utterance to which the discourse marker is attached is to be interpreted. This is the best studied group of pragmatic markers (see e.g. Schiffrin 1987, Blakemore 2002, Fraser 1999 among others; for Bulgarian see Tisheva, Hauge 2001, 2002).

Communication as an interactive process requires speakers to show different type of communicative knowledge. "Discourse markers are one set of linguistic items that function in cognitive, expressive, social, and textual domains." (Schiffrin 2001: 54). Heeman and Allen (1999) state that the utterance units are the building blocks of spoken dialogues and discourse markers operate at this level to relate the current utterance to the previous or forthcoming context. Here we will define these items as linguistic devise that speakers use to signal how the upcoming unit of speech or text relates to the current discourse state. Although the discourse markers in contrast to the other pragmatic markers do not contribute to the representative sentence meaning they are so prominent in dialogues so that they could be valuable source of information for understanding the utterances that they introduce.

The discourse markers identification could be done at a very early stage of annotation, even to be incorporated to the POS-tagging of a dialogue corpus (see Heeman and Allen 1999). The complications and possible ambiguities are due to the fact that by origin the discourse markers come from different parts of speech – verbs, conjunctions, particles. Multiword markers – like *Разбираш ли*, imperatives like *gledaj, viz* (See! Look!), used to introduce an utterance possibly will get then more than one interpretations. The complication here is also due to the fact that the discourse markers tend to be used to introduce a new utterance, or can be an utterance all to themselves, or can be used as part of editing term of a speech repair. Thus, the problem of identify-

ing the discourse markers also needs to be addressed with the segmentation and speech-repair problem.

Discourse markers which originate from the groups of particles (like English *well,* Bulgarian *nali*) or conjunctions (Bulgarian *ami, ama, znači,*) are ambiguous as to whether they are being used as discourse markers or not. In example (14) the use of the adversative conjunction *znači* as a discourse marker is illustrated, while in (15) the same lexical item functions to connect the parts of the complex sentence:

(14) priznaxme si znači greškata a posle si prostihme
We confessed our mistake, and after that we forgave each other.

(15) Pari njama znači i кredit njama
No money, so no credit.

Although the discourse markers tend to occupy the beginning of the utterance, in the case of particles and conjunctions this is not always observed. The position of *znaci* in the mid-part of the sentence in (14) is significant for its function on the discourse level. For contrast, in (15) it is used to introduce the second part of the complex sentence.

Similarly, the particle *nali* can be used as a discourse marker inside the utterance, as shown in (16), but it can be also used as a question particle, to mark the illocutionary force of the sentence, as this is illustrated by (17):

(16) nači tazi arxitektkata mi dade tozi varaint /
kato / nali / nejno predloženie
The architect gave me this idea as her suggestion.

(17) Az nali te popitax za tova?
Didn't I ask you about this?

The phenomena of spoken dialogues discussed here cannot be resolved without recourse to the syntactic information. Enriching the annotation with a set of specific features will benefit not only the discourse analyses, but will also contribute to the effective use of the corpus as representative data base for spoken Bulgarian.

IV. Conclusion

The present paper has provided an overview of some main phonological, grammatical and discourse features of spoke (colloquial) Bulgarian relevant for its representation in data bases and corpora. In order to support and benefit future work on coding/processing language data on different levels we discussed the ways of possible annotation of these features. We consider the data structure and appearance to be the most significant features, which will distinguish the corpus of spoken Bulgarian from those of written or standard texts. It is hoped that the choice between original, normalized or regularized forms of transcripts, the careful selection of annotated features of verbal and non-verbal communicative events could make the data more theory-neutral. The flexibility of representation and the alternatives in displaying the information needed for different research purposes are the goals to achieve in the future.

References

Blakemore, D. 2002. *Linguistic Meaning and Relevance: The Semantics and Pragmatics of Discourse Markers.* Cambridge: Cambridge University Press.

Edwards, J. A. 2001. 'The Transcription of Discourse', in *The Handbook of Discourse Analysis.* Ed. D. Schiffrin, D. Tannen & H. Hamilton. 54–75. Oxford: Blackwell, 321–347.

Fraser, B. 1996. 'Pragmatic markers', *Pragmatics.* Vol. 6, No 2, 167–190.

Fraser, B. 1999. 'What are discourse markers?', *Journal of Pragmatics*, 31.

Heeman, P. A., J. F. Allen. 1997. 'Intonational boundaries, speech repairs, and discourse markers: Modeling spoken dialog', in *Proceedings of the 35*th *Annual Meeting of the Association for Computational Linguistics*, 254–261.

Heeman, P. A., J. F.Allen. 1999. 'Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialog', *Computational Linguistics*, 25(4), 527–572.

Krapova, I., Tisheva, Y. 2006. 'Clitic reduplication Structures in the Bulgarian Dialects', in *Diahronija in sinhronija v dalektoloskih raziskavah.* Maribor: Zora, 41, 415–421.

Schiffrin, D. 1987. *Discourse Markers.* (Studies in International Sociolinguistics, 5). Cambridge: Cambridge University Press.

Schiffrin, D. 2001. 'Discourse Markers: Language, Meaning, and Context', in *The Handbook of Discourse Analysis.* Ed. D. Schiffrin, D. Tannen & H. Hamilton. 54–75. Oxford: Blackwell, 54–75.

Stubbs, M. 2001. 'Computer-assisted Text and Corpus Analysis: Lexical Cohesion and Communicative Competence', in *The Handbook of Discourse Analysis.* Ed. D. Schiffrin, D. Tannen & H. Hamilton. 54–75. Oxford: Blackwell, 305–319.

Тишева, Й., Х. Хауге. 2001. 'Съюзи за противопоставяне в ролята на прагматични частици', in *Проблеми на българската разговорна реч*, т. 5. Велико Търново, 242–252.

Тишева, Й., Х. Хауге. 2002. 'Заемане на прагматични частици', in *Билингвизъм и диглосия. Проблеми на социолингвистиката*, т. 7. София, 10–17.