**Computer Applications in Slavic Studies**

Proceedings of *Azbuky.Net*

International Conference and Workshop

24–27 October 2005, Sofia, Bulgaria

# *Computer Processing of Modern Languages*

## Encoding Bulgarian Colloquial Speech Using the TEI Specification

*Atanas Atanasov*

*(Faculty of Slavic Studies, Department of Bulgarian Language*

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

**Abstract**: This paper presents a model for creation of a corpus of transcribed Bulgarian colloquial speech. The main goal is to show how the TEI specification is used for resolving some problems in XML encoding of spontaneous speech. The first step is to determine the scope of the description and – respectively – the elements, included in the DTD, and their attributes. In the header of the XML documents, as usual, there is the meta-information (which includes text information, description of the participants, duration of the record etc.). In the body, for now, only a partial syntactic annotation is presented. The description of some pragmatic features such as the illocutionary force of the utterance (interrogative, exclamatory etc.), extra-linguistic phenomena such as facial expressions, gestures, pauses, etc. is also discussed. We have attempted to manage with some difficulties, for instance overlapping of the utterances, incomplete sentences, etc.

The main goal of our efforts is to create a syntactically annotated corpus of spoken Bulgarian in XML format using the TEI specification. The first question is what kind of features should be included in the annotation. The data is divided in two primary parts – metadata (including text information, description of the participants, duration of the record etc.) and information, describing some linguistic phenomena (at different levels: phonology, morphosyntax,

pragmatic and discourse features). For the creation of the DTD the "TEI Pizza Chef" (http://www.tei-c.org/pizza.html) was used.

The basic document structure contains the following elements:

```
- <TEI.2>
  - <teiHeader>
    + <fileDesc></fileDesc>
    + <profileDesc></profileDesc>
    + <revisionDesc></revisionDesc>
    </teiHeader>
  - <text>
    + <body></body>
    </text>
  </TEI.2>
```

In the header (where the metadata is stored) there are three main elements: <fileDesc>, <profileDesc> and <revisionDesc>. The first one contains information about the file (title, author, record extent, publication date and place, source description, etc.). The second one gives information about the text (language, date and place of recording, participants' description, text features, etc.). The third one contains information about the verification of the transcription.

The body of the document contains the corpus itself and it will be discussed more precisely, due to the fact that the main difficulties appear here. The text is constructed by utterances and each utterance is enclosed by the element <u>. It has an attribute "*who*", which supplies an identifier for the speaker and its value corresponds to the value of the "*id*" in the element <person> (in the participant description in the header). The utterance may be continuous or could be interrupted by pauses. The element <pause> is empty, and has an attribute "*dur*", which indicates the approximate length of this element in time – short, medium or long (the "\" character indicates the stress):

```
- <u who="AA">
    \да и в чет\вьртък сутрин\ть \ставаш и си \тръгваш
    <pause dur="medium"/>
    та\ка че \е на \другия \ден
  </u>
```





Further in this paper some linguistic end extra-linguistic phenomena will be discussed. We will show the way we handle some difficulties in their annotation.

**Incomplete utterances**

Sometimes the speaker is interrupted by another participant in the conversation. In these cases there are three possibilities:

```
- <u who="BA">
    \ох
    <pause dur="short"/>
    \ужас
    <pause dur="short"/>
    <seg part="I">о\тиваме \е \са в \петък пре\ди \да </seg>
  </u>
  <u who="AA"> на \следващия \ден ? </u>
```

This example presents an utterance, which is "partial" – it is incomplete (unfinished). The speaker has started talking, but he has been interrupted, hence only the initial part of the utterance is available. That is why the element <seg> is needed – it marks certain fragment of the utterance and its attribute "*part*" shows which part exactly is dropped and which is available. For example <seg part="I"> ... </seg> means that the particular fragment is "initial", i.e. it's not finished; <seg part="F"> ... </seg> means that the particular fragment is "final", i.e. the initial part is skipped, like in this case:

```
  <u who="AA"> е как\во \нищо ? </u>
- <u who="BA">
    <vocal desc="ъ"/>
    <seg part="F">и я зап\лашвал с то\ва </seg>
    <pause dur="short"/>
    \тя \да си о\тиде на \костинброд и нак\рая \тя си о\тиде на \костинброд
  </u>
```

The next example illustrates a more complicated type of partial utterances: the first speaker starts talking, the second speaker interrupts him, and after that the first one continues his sentence:

```
    <pause dur="medium"/>
    <seg part="I"> \снощи </seg>
  </u>
- <u who="BA">
    <seg part="M"> в \осем </seg>
  </u>
- <u who="AA">
    <seg part="F">в \девет си си \легнала </seg>
    <pause dur="medium"/>
```

The second speaker's utterance is marked as a segment, whose attribute's value is "M" (a medial part of an incomplete segment). It means that neither the initial, nor the final parts of the utterance are available. Actually, in this particular case, they both are in the first's speaker utterances, and the second speaker's utterance is inserted between them, so the three segments combined together form one syntactic unit.

**False start**

With this term we indicate the kind of incomplete utterance – when the speaker starts talking, but for some reason he or she decides to change the subject, so the fragment in the element <seg> is unfinished (that is why its attribute's value is always "I" – initial):

```
  </u>
- <u who="AA">
    сак\сията \е
    <seg part="I">със </seg>
    стра\хотна
    <pause dur="medium"/>
    \имам \нарове
  </u>
```

**Speech repair**

The speech repair is a phenomenon, similar to the false start. The difference is that here the subject (the theme of conversation) is not changed; the speaker goes back to a previous part of his own utterance to make it more precise or to correct it:





```
    <seg part="I"> \чак нак\рая се запоз\нава </seg>
    <pause dur="short"/>
    \деси \най - нак\рая се запоз\нава със \нея
    <pause dur="short"/>
```

**Incomplete words**

A speech repair, a false start, or just a damaged record sometimes could be also a reason for incomplete words in the corpus. As they do not form syntactic units, the element <seg> is not appropriate for their annotation; the element <w> is used instead. When the reason for the interruption of the word is pragmatic (not a damaged tape for example), the only possible value for the attribute "*part*" is "I" (initial):

```
- <seg part="I">
    на \другия \ден теле\фона му е \цял
    <w part="I">д </w>
  </seg>
  \те не из\лизат вечер\та на \другия \ден му \е \цял \ден теле\фона изк\лючен
  <pause dur="short"/>
```

**Unclear text**

When somewhere in the record a unclear part is detected (a damaged tape or a loud noise for example) the element <gap> is used. Its attribute "*reason*" explains why there is a gap in the utterance – a unclear part of the record, a noise etc.

```
- <u who="BA">
    ше му се о\бада и ше му \кажа
    <gap reason="unclear"/>
  </u>
- <u who="AA">
    \даже \тряше \по - \раничко \да му \звъннеш
    <pause dur="short"/>
    спо\ред \мен
  </u>
```

**Vocalized semi-lexical**

In the TEI specification the element <vocal> indicates any vocalized but not necessarily lexical phenomenon, for example voiced pauses, non-lexical backchannels, etc. The argument "*desc*" explains the exact meaning of the vocalized semi-lexical:

```
- <u who="AA">
    \ти \кво ше \режеш се\га
    <pause dur="short"/>
    <vocal desc="м - въпросително"/>
  </u>
```

**Non-vocalized quasi-lexical**

This term is used for the description of not vocalized communicative phenomena like a gesture and a facial expression. The argument "*desc*" supplies a representation of the phenomenon:

```
- <u who="OB">
    \слави \трифонов ха\ресваш ли го ?
    <pause dur="short"/>
  </u>
- <u who="BA">
    <kinesic desc="изкривява устни в израз на отрицание"/>
  </u>
```

**Overlapping**

Sometimes different parts of the utterances produced by the speakers could overlap. The simultaneous talk of two (or more) speakers is one of the main difficulties in annotating the transcribed texts. Here a very convenient decision is chosen – the element <anchor> gives an opportunity to put timelines in the text:





```
- <u who="AA">
    \ми \да \де \именно
    <pause dur="short"/>
    <anchor id="t09"/>
    \те от то\ва се полу\чават
    <anchor id="t10"/>
  </u>
- <u who="BA">
    <anchor synch="t09"/>
    не \знам от \кво
    <anchor synch="t10"/>
    се по\лучиха \теа \мравки ?
  </u>
```

This way not only the most frequent case when the end of the first utterance overlaps with the start of the next utterance (like in the example above), but much more complicated situations could be presented:

```
- <u who="AA">
    \дай \дай
    <anchor id="t19"/>
    \аз ше си ги
    <anchor id="t20"/>
    на\режа
  </u>
- <u who="BA">
    <anchor synch="t19"/>
    \по - \едро
    <anchor synch="t20"/>
  </u>
```

In this case the whole utterance of the second speaker overlaps with the middle part of the first speaker's utterance.

**Illocutionary force**

When the utterance is not in the indicative mood its illocutionary force is marked too (when it is interrogative or imperative for example):

```
- <u who="OB">
    <seg type="interrogative">\да \не \са със \заешки \пух </seg>
  </u>
```

This paper has provided an overview of the main difficulties we met in our attempts to create an annotated corpus of colloquial Bulgarian speech. The main goal to achieve in the future is to encode the transcribed texts more precisely, providing more detailed information about the morphological characteristics of the words and about the syntactic units (the constituent groups – NP, VP, AP, etc.). It is also important to mark the main features which distinguish the linguistic data in the corpus of spoken Bulgarian from those of written or standard texts.

**References**

http://bgspeech.net
http://bultreebank.org
http://www.tei-c.org