BTB-TR04: BulTreeBank Morphosyntactic Annotation of Bulgarian Texts*

Kiril Simov, Petya Osenova
BulTreeBank Project
http://www.BulTreeBank.org
Linguistic Modelling Laboratory, Bulgarian Academy of Sciences
Acad. G. Bonchev St. 25A, 1113 Sofia, Bulgaria
kivs@bultreebank.org, petya@bultreebank.org

BulTreeBank Technical Report BTB-TR04

05.04.2004

Abstract

This document presents the guidelines for the morphosyntactic annotation of Bulgarian texts within the BulTreeBank Project. The morphosyntactic annotation is based on BTB-TS tagset decribed in [Simov et. al. 2004]. Here we discuss the following topics: token classification; token level annotation with morphological lexicon, gazetteers, finite state grammars; morphosyntactic ambiguities; semi-automatic disambiguation; and validation.

1 Introduction

This document focuses on the basic decisions taken during the creation of the BulTreeBank morpho-syntactic corpus. It describes the classification of the tokens in the text. Then, the token level annotation for each type of tokens is presented. In the process of token level annotation we have consulted the following knowledge-based resources: *the morphological lexicon, gazetteers, finite state grammars.* Next, we describe ambiguities at two levels: among different kinds of tokens and within a given class of tokens. Note that we have focused on the most frequent ambiguities, because they are assumed to have greater impact on the analyzed data. For each type of ambiguity we give rules and hints for disambiguation with different degree of certainty. Some of the 100 % sure rules are implemented as constraints in CLaRK System and they are used for automatic disambiguation. For the rest of the cases semi-automatic disambiguation is applied, i.e. some expert inspection is always involved. The semi-automatic disambiguation is also supported by a set of constraints. The last section discusses the validation process.

2 Token Classification

The annotation of tokens in the BulTreeBank morpho-syntactic corpus is as follows:

*The work reported here is done within the BulTreeBank project. The project is funded by the Volkswagen Stiftung, Federal Republic of Germany under the Programme *"Cooperation with Natural and Engineering Scientists in Central and Eastern Europe"* contract I/76 887.

2.1 Common Words

They include the common nouns, verbs, adjectives and other parts of speech. As it was mentioned in BTB-TR02, these words are written as a rule in small letters inside the sentences and with a capital letter in a sentence initial position.

We annotate each common word within the following XML element:

```
<w aa="MSD_List" ana="MSD">wordform</w>
```

Here `MSD_List` is a list of tags from BTB-TS tagset separated by semicolons. These tags represent all possible morpho-syntactic features of the wordform. The value of attribute `ana` is one of the members of `MSD_List`. It is the correct morpho-syntactic features for the wordform in the given context of its usage. Two additional attributes are discussed below. Here is one sentence annotated in this way.

```
<text>Подсъдимият не разбираше нищо от речта на защитника си.</text>
<s>
  <w aa="Amsf;Ncmsf" ana="Ncmsf">Подсъдимият</w>
  <w aa="Tn" ana="Tn">не</w>
  <w aa="Vpitf-m2s;Vpitf-m3s" ana="Vpitf-m3s">разбираше</w>
  <w aa="Pne-os-ni" ana="Pne-os-ni">нищо</w>
  <w aa="R" ana="R">от</w>
  <w aa="Ncfsd" ana="Ncfsd">речта</w>
  <w aa="R;Te" ana="R">на</w>
  <w aa="Ncmsh" ana="Ncmsh">защитника</w>
  <w aa="Ncnsi;Ppxtd;Psxto;Vxitf-r2s" ana="Psxto">си</w>
  <pt>.</pt>
</s>
```

Some wordforms consist of more than one token. We divide these wordforms into two categories: with a **fixed positional structure** — some pronouns, conjunctions, prepositions, adverbs, numerals; with a **varying positional structure** — analytical verb forms, verb-clitic forms, nominal-clitic forms. The former one is recognized at the token annotation level. It is annotated similarly to the one-token words, but the XML element has the tag `mw` instead of `w`:

```
<mw aa="MSD_List" ana="MSD">wordform</mw>
```

The meaning of the values of the two attributes is the same as above. The multi-wordform contains intervals between the tokens. Here is an example:

```
<text>Съвсем загуби власт над себе си.</text>
<s>
  <w aa="Dq" ana="Dq">Съвсем</w>
  <w aa="Ncfpi;Vpptf-o2s;Vpptf-o3s;Vpptf-r3s;Vpptz--2s" ana="Vpptf-o3s">загуби</w>
  <w aa="Ncfsi" ana="Ncfsi">власт</w>
  <w aa="R" ana="R">над</w>
  <mw aa="Ppxla" ana="Ppxla">себе си</mw>
  <pt>.</pt>
</s>
```

The latter kind is recognized via regular grammars at the next level of annotation:

```
<text>Един непознат човек ме спря.</text>
```

```
<s>
  <w aa="Mcmsi;Pfe-os-mi" ana="Pfe-os-mi">Един</w>
  <w aa="Amsi;Ncmsi" ana="Amsi">непознат</w>
  <w aa="Ncmsi" ana="Ncmsi">човек</w>
  <V>
  <w aa="Ppetas1" ana="Ppetas1">ме</w>
  <w aa="Vpptf-o2s;Vpptf-o3s" ana="Vpptf-o3s">спря</w>
  </V>
  <pt>.</pt>
</s>
```

The division between multi-token wordforms with fixed and varying structure is governed by two criteria: *compositionality* and *adjacency*. According to the first criterion: we consider the following multi-token syntactic elements: compositional phrases which we discuss in [Osenova and Simov 2004], non-compositional phrases also discussed in [Osenova and Simov 2004], compositional and non-compositional multi-token wordforms. Compositional multi-token wordforms consist of lexical elements only and cannot incorporate potential phrasal elements. On the basis of the last requirement we have reanalyzed some of the traditional multi-token wordforms as phrases. The most significant multi-token wordforms, which have been re-analyzed in this way, are some analytical verb forms with the verb ща ('sta' will) and some compound conjunctions consisting of a preposition and the particle да ('da' to). The non-compositional multi-token wordforms consist of tokens which have their own interpretation, but the whole meaning is not the sum of these distinct interpretations ('може би', maybe; 'който и да е', whoever).

The second criterion plays a crucial role for the phrase-like or word-like division. For example, the analytical verb forms headed by the past form of the verb ща ('sta' will) plus da-construction are considered phrases, because other elements can be inserted among them: щях ли аз да дойда?, 'would interrogative-particle I to come?', would I come?

Old usage. There are some old words that are used in contemporary texts. These words in the texts are divided into two groups: (1) homonymic wordforms — the old wordform has the same spelling as some contemporary wordform; and (2) old wordform differs from all contemporary wordforms. The wordforms of the first kind usually differ from the contemporary forms in their grammatical features. In order to make difference between an old wordform and contemporary ones we use the attribute `usage` with value `old`. In these cases the wordform is annotated only with the morpho-syntactic information as an old wordform. When the same wordform is used in its contemporary usage, it is annotated with its contemporary morpho-syntactic features. The second kind of old wordforms are annotated with their morpho-syntactic features, but they are not marked as old usage, because: (1) they have this information in the tag, or (2) the tag does not coincide with any contemporary usage and thus the old usage marking becomes irrelevant.

Here is an example for the first kind:

```
<text>Старата не ще изтрай.</text>
<s>
  <w aa="Afsd" ana="Afsd">Старата</w>
  <w aa="Tn" ana="Tn">не</w>
  <w aa="Tx;Vpitf-r3s" ana="Tx">ще</w>
  <w aa="Vpptf-r3s" ana="Vpptf-r3s" usage="old">изтрай</w>
  <pt>.</pt>
</s>
```

In this example the wordform изтрай in its old usage coincides with the contemporary form for imperative.

Here is an example of the second kind:

```
<text>Обвиняемий, пристъпете по-близо.</text>
```

```
<s>
  <w aa="Ncms-v" ana="Ncms-v">Обвиняемий</w>
  <pt>,</pt>
  <w aa="Vppiz--2p" ana="Vppiz--2p">пристъпете</w>
  <w aa="Ansi;Dl" ana="Dl">по-близко</w>
  <pt>.</pt>
</s>
```

In this example the wordform Обвиняемий is an old form for vocative, but it does not coincide with a contemporary form and thus it does not need to be specially annotated.

Special symbols as words. There are some special symbols that are frequently used instead of common words. The most frequent symbols of this kind are: *figures, per cent sign, dollar sign, colon, slash.* Some of them are uniquely defined with respect to their usage, but others are not and therefore, they allow for more interpretations. In order to make such an usage clear, we introduce the attribute `exp` for the elements `<w>` and `<mw>`. The value of this attribute is the actual wordform substituted by the symbol(s). The values for figures are given only if necessary.

Example:

```
<text>Инвалидите плащат 50 % от дължимия налог.</text>
<s>
  <w aa="Ncmpd" ana="Ncmpd">Инвалидите</w>
  <w aa="Vpitf-r3p" ana="Vpitf-r3p">плащат</w></V>
  <w aa="Mc--i;Mo-pi;Mofsi;Momsi;Monsi" ana="Mc--i">50</w>
  <w aa="Ncmsh;Ncmsi;Ncmt" ana="Ncmt" exp="процента">%</w>
  <w aa="R" ana="R">от</w>
  <w aa="Amsh" ana="Amsh">дължимия</w>
  <w aa="Ncmsi" ana="Ncmsi">налог</w>
  <pt>.</pt>
</s>
```

Here are the definitions of the two elements and their attributes from the BulTreeBank DTD:

```
<!ELEMENT w (#PCDATA) >
<!ATTLIST w
    aa    CDATA #REQUIRED
    ana   CDATA #REQUIRED
    exp   CDATA #IMPLIED
    usage CDATA #IMPLIED
>

<!ELEMENT mw (#PCDATA) >
<!ATTLIST mw
    aa    CDATA #REQUIRED
    ana   CDATA #REQUIRED
    exp   CDATA #IMPLIED
    usage CDATA #IMPLIED
>
```

2.2 Proper Names

We have annotated the proper names at both levels: token and phrasal. Each name is categorized ontologically as a person, location, organization or other entity. Each token level name is annotated as follows:

```
<name aa="MSD_List" ana="MSD" cat="Cat" sort="Sort">Name</name>
```

Here `MSD_List` is a list of tags from BTB-TS tagset separated by semicolons. These tags represent all possible morpho-syntactic features of the name. The value of attribute `ana` is one of the members of `MSD_List`. It is the correct morpho-syntactic features for the name in the given context of its usage. The value of the attribute `sort` is one or several ontological categories. The possible values are: `NE-Pers` for persons, `NE-Loc` for locations, `NE-Org` for organizations, and `NE-Other` for other entities. The attribute `cat` determines the category of the name. This is necessary because some very popular names consisting of more that one token are annotated in one `name` tag, but later are considered as phrases. Such names include: Стара планина ('Stara planina' the Balkan), Брус Уилис ('Brus Uilis' Bruce Willis). The two values for the attribute `cat` are: `lex` for lexical names and `NPA` for phrasal names.

Along with the usual names we also annotate as names all kinds of unique segments of texts which designate unique entities and which are not common words or abbreviations. Such text fragments can be e-mail addresses, URLs, IP addresses, plate signs, letter names etc.

Here are two example sentences with names:

```
<text>Гроздан се дразнеше все повече.</text>
<s>
  <name aa="Npmsi" ana="Npmsi" cat="lex" sort="NE-Pers">Гроздан</name>
  <w aa="Ppxta" ana="Ppxta">се</w>
  <w aa="Vpitf-m2s;Vpitf-m3s" ana="Vpitf-m3s">дразнеше</w>
  <w aa="Dq" ana="Dq">все</w>
  <w aa="Mc--i" ana="Mc--i">повече</w>
  <pt>.</pt>
</s>

<text>Бил Клинтън ще се кандидатира за кмет на Ню Йорк?</text>
<s>
  <name aa="Npmsi" ana="Npmsi" cat="NPA" sort="NE-Pers">Бил Клинтън</name>
  <w aa="Tx;Vpitf-r3s" ana="Tx">ще</w>
  <w aa="Ppxta" ana="Ppxta">се</w>
  <w aa="Vpiif-o2s;Vpiif-o3s;Vpiif-r3s;Vppif-o2s;Vppif-o3s;Vppif-r3s"
     ana="Vppiif-r3s">кандидатира</w>
  <w aa="R" ana="R">за</w>
  <w aa="Ncmsi" ana="Ncmsi">кмет</w>
  <w aa="R;Te" ana="R">на</w>
  <name aa="Npmsi" ana="Npmsi" cat="lex" sort="NE-Loc">Ню Йорк</name>
  <pt>?</pt>
 </s>
```

In this example the name Бил Клинтън ('Bill Clinton') is recognized as one element although both names Бил and Клинтън are present in the gazetteer. Processing some of the names in this way helps us to facilitate the later processing of the data in the treebank. If someone needs these two names to be annotated separately, it is an easy task to separate them. Note that the value of the attribute `cat` — `NPA` is a tag from the annotation scheme for such names. It means that the noun phrase consists of a head and an adjunct. The `lex` value for the other names in the examples means that the immediate syntactic tag for them will depend on their morpho-syntactic tags.

Sometimes the same name can be used to designate several entities. Very often person's names are used for locations, companies, sport teams etc. In this case we annotate the name with the actual category, or with the basic entity, if no other clues for the actual usage are available. At the level of syntax we annotate more complex names.

It is well known that frequently names coincide with common words or abbreviations. In order to distinguish such names we annotate them with an additional attribute `amb` with two possible values: `on`

and `off`. If the value is `"on"` then the name could be also a common word (or an abbreviation) and thus we need to be more careful in distinguishing between a name or a common word. Also, this value is applicable to the common words when they are used as names. In the last case the morpho-syntactic tags are corresponding to the tags for names.

Here is an example:

```
<text>Новият игрален филм е режисиран от Иглика Трифонова.</text>
<s>
  <w aa="Amsf" ana="Amsf">Новият</w>
  <w aa="Amsi" ana="Amsi">игрален</w>
  <w aa="Ncmsi" ana="Ncmsi">филм</w>
  <w aa="I;Te;Vxitf-r3s" ana="Vxitf-r3s">е</w>
  <w aa="Vpitcv--smi;Vpptcv--smi" ana="Vpptcv--smi">режисиран</w>
  <w aa="R" ana="R">от</w>
  <name aa="Npfsi" amb="on" ana="Npfsi" cat="lex" sort="PersNE">Иглика</name>
  <name aa="Hfsi" ana="Hfsi" cat="lex" sort="PersNE">Трифонова</name>
  <pt>.</pt>
</s>
```

In this example the first name Иглика is also a common noun for a flower and this is why it is annotated with the attribute `amb` with value `on`. The capital letter can be a disambiguation feature in the middle of a sentences, but not at the beginning of it.

Here is the definition of the element and its attributes from the BulTreeBank DTD:

```
<!ELEMENT name (#PCDATA) >
<!ATTLIST name
    aa CDATA #REQUIRED
    amb (on|off) "off"
    ana CDATA #REQUIRED
    cat CDATA #REQUIRED
    sort (NE-Pers|NE-Loc|NE-Org|NE-Other) #REQUIRED
>
```

2.3 Abbreviations

The abbreviations are of two types: *acronyms* and *contractions*. The former usually designate named entities, the latter correspond to common words or short phrases. When an abbreviation designates a named entity it is categorized ontologically as a person, location, organization or other entity. Each abbreviation is annotated as follows:

```
<abbr aa="MSD_List" ana="MSD" cat="Cat" sort="Sort" type="Type" exp="Exp">Abbr</abbr>
```

Here `MSD_List` is a list of tags from BTB-TS tagset separated by semicolons. These tags represent all possible morpho-syntactic features of the name. The value of attribute `ana` is one of the members of `MSD_List`. It is the correct morpho-syntactic features for the name in the given context of its usage. The value of the attribute `sort` is one or several ontological categories. The possible values are: `NE-Pers` for persons, `NE-Loc` for locations, `NE-Org` for organizations, and `NE-Other` for other entities. Because some of the abbreviations correspond to common words or phrases, the attribute `sort` receives a value `common` for such abbreviations. Usually, it is presented for acronyms. The attribute `cat` determines the category of the abbreviation. The possible values for the attribute `cat` include `lex` for lexical abbreviations and a wider range of phrasal values in comparison with the names. This is due to the fact that abbreviations can correspond to a wider range of phrases. The attribute `type` determines the kind of the abbreviation.

The possible values are: `acr` for acronyms and `contr` for contractions. Each abbreviation has at least one word or expression it stands for. This word or expression is given as a value for the attribute `exp`. This value represents the actual expression in the context.

Here are some examples of sentences containing abbreviations:

```
<text>Лидерите подготвят нова декларация за външната политика на ЕС.</text>
<s>
   <w aa="Ncmpd" ana="Ncmpd">Лидерите</w>
   <w aa="Vpitf-r3p;Vpptf-r3p" ana="Vpitf-r3p">подготвят</w>
   <w aa="Afsi" ana="Afsi">нова</w>
   <w aa="Ncfsi" ana="Ncfsi">декларация</w>
   <w aa="R" ana="R">за</w>
   <w aa="Afsd" ana="Afsd">външната</w>
   <w aa="Ncfsi;Ncmsh" ana="Ncfsi">политика</w>
   <w aa="R;Te" ana="R">на</w>
   <abbr aa="Npmsi" ana="Npmsi" cat="NPA"
         exp="Европейски съюз" sort="NE-Loc" type="acr">ЕС</abbr>
   <pt>.</pt>
</s>

<text>Важно изискване е оборотът да е под 75 000 лв.</text>
<s>
   <w aa="Ansi;Dm" ana="Ansi">Важно</w>
   <w aa="Ncnsi" ana="Ncnsi">изискване</w>
   <w aa="I;Te;Vxitf-r3s" ana="Vxitf-r3s">е</w>
   <w aa="Ncmsf" ana="Ncmsf">оборотът</w>
   <w aa="Ta;Tx" ana="Tx">да</w>
   <w aa="I;Te;Vxitf-r3s" ana="Vxitf-r3s">е</w>
   <w aa="Ncmsi;R" ana="R">под</w>
   <w aa="Mc--i;Mo-pi;Mofsi;Momsi;Monsi" ana="Mc--i">75 000</w>
   <abbr aa="Ncmsh;Ncmsi;Ncmt" ana="Ncmt" cat="lex"
         exp="лева" type="contr">лв.</abbr>
   <pt>.</pt>
</s>
```

Similarly to names, abbreviations frequently coincide with common words or names. In order to distinguish such abbreviations we annotate them with an additional attribute `amb` with two possible values: `on` and `off`. If the value is `"on"` then the abbreviation could be also a common word (or a name) and thus we need to be more careful in distinguishing between an abbreviation or a common word. This is especially important for the recognition of the sentence boundaries.

Here is the definition of the element and its attributes from the BulTreeBank DTD:

```
<!ELEMENT abbr(#PCDATA)>
<!ATTLIST abbr
    aa   CDATA #REQUIRED
    amb  (on|off) "off"
    ana  CDATA #REQUIRED
    cat  CDATA #REQUIRED
    exp  CDATA #REQUIRED
    sort (NE-Pers|NE-Loc|NE-Org|NE-Other|common) "common"
    type CDATA #REQUIRED
>
```

2.4 Foreign Tokens

Often in Bulgarian texts there are tokens written in Latin alphabet. Such tokens can be isolated words, well known abbreviations or names and complete phrases. Thus we classify them according to these categories. If the token is a well known word, abbreviation or name we use the above elements: `<w>`, `<name>`, `<abbr>` with appropriate attribute's values. If there is a phrase or a non-well known word (usually accompanied by an explanation in Bulgarian) we use the element `<foreign>`:

```
<foreign aa="MSD_List" ana="MSD" cat="Cat" lang="Lang">Text</foreign>
```

Here `MSD_List` is a list of tags from BTB-TS tagset separated by semicolons. These tags represent all possible morpho-syntactic features of the token. The value of attribute `ana` is one of the members of `MSD_List`. It is the correct morpho-syntactic features for the token in the given context of its usage. Usually the morpho-syntactic features for such phrases are underspecified and include only the part of speech information. The attribute `cat` determines the category of the phrase. The value `lex` stands for lexical tokens and then the syntactic category is determined on the basis of the morpho-syntactic features. Otherwise the value determines the phrasal category and it is taken from the annotation scheme. The attribute `lang` determines the original language of the phrase only if it can be identified non-problematically.

Sometimes the foreign material is transliterated with Cyrillic letters. Also, we use the same element to annotate such phrases.

Here is an examples:

```
<text>Ако използваме англосаксонското понятие за checks and balances -
          уравновесяване и противопоставяне, този баланс...</text>
<s>
   <w aa="Cs" ana="Cs">Ако</w>
   <w aa="Vpitf-r1p" ana="Vpitf-r1p">използваме</w>
   <w aa="Ansd" ana="Ansd">англосаксонското</w>
   <w aa="Ncnsi" ana="Ncnsi">понятие</w>
   <w aa="R" ana="R">за</w>
   <foreign aa="N" ana="N" cat="NPA" lang="eng">checks and balances</foreign>
   <pt>-</pt>
   <w aa="Ncnsi" ana="Ncnsi">уравновесяване</w>
   <w aa="Cp" ana="Cp">и</w>
   <w aa="Ncnsi" ana="Ncnsi">противопоставяне</w>
   <pt>,</pt>
   <w aa="Pde-os-m" ana="Pde-os-m">този</w>
   <w aa="Ncmsi" ana="Ncmsi">баланс</w>
   <pt>...</pt>
</s>
```

Here is the definition of the element and its attributes from the BulTreeBank DTD:

```
<!ELEMENT foreign (#PCDATA) >
<!ATTLIST foreign
    aa   CDATA #REQUIRED
    ana  CDATA #REQUIRED
    cat  CDATA #REQUIRED
    lang CDATA #IMPLIED
>
```

2.5 Figures

The figures are annotated as common words with the element <w> with corresponding morpho-syntactic features. Here we include all arabic, roman figures and fractions. A figure can contain a comma (or a dot), a slash or a space. For example: <w>1999</w>, <w>19,32</w>, <w>0.03</w>, <w>1/8</w>, <w>75 000</w>, <w>XIX</w>. Here we omitted the attributes. Some figures in the text are parts of other tokens and they are segmented together with the rest of those tokens. Examples of this model are: МиГ-29, B2B, etc.

The value of aa attribute is determined on the basis of the actual number. Such figures are more frequently cardinal, but sometimes could be ordinals, although ordinals are usually written with ending showing the grammatical features like 1-ата, 7-ят. But sometimes such endings are not presented as in **1991 година**. This is why we include both kind of tags in the value of attribute aa. The general rules are as follows:

- **Integers 1 and -1**.

  In these cases the possible grammatical features are these of the cardinal number one (един (Mcmsi), една (Mcfsi), едно (Mcnsi)) and the ordinal number first (пръв (Momsi), първи (Momsi), първа (Mofsi), първо (Monsi), първи (Mo-pi)). Here are some examples:

```
<w aa="Mcfsi;Mcmsi;Mcnsi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mcmsi">1</w>
<w aa="Ncmsi" ana="Ncmsi">грам</w>

<w aa="Mcfsi;Mcmsi;Mcnsi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mcfsi">1</w>
<w aa="Ncfsi" ana="Ncfsi">чаша</w>

<w aa="Mcfsi;Mcmsi;Mcnsi;Mo-pi;Mofsi;Momsi;Monsi" ana="Momsi">1</w>
<w aa="Ncmsi" ana="Ncmsi">век</w>

<w aa="Mcfsi;Mcmsi;Mcnsi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mofsi">1</w>
<w aa="Ncfsi" ana="Ncfsi">гимназия</w>
```

- **Integers ending in 1** (21, 31, ... without numerals ending in 11).

  In these cases the possible grammatical features are these of the cardinal number three, or four etc. (три (Mc-pi)) and the ordinal number first (пръв (Momsi), първи (Momsi), първа (Mofsi), първо (Monsi), първи (Mo-pi)). Note that in the non-ordinal readings the whole integer gets the plural grammatical feature of number before 1, because the integer governs a plural or count form of a noun. Here are some examples:

```
<w aa="Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mc-pi">31</w>
<w aa="Ncmsh;Ncmt" ana="Ncmt">грама</w>

<w aa="Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mc-pi">41</w>
<w aa="Ncfpi" ana="Ncfpi">чаши</w>

<w aa="Mcfsi;Mcmsi;Mcnsi;Mo-pi;Mofsi;Momsi;Monsi" ana="Momsi">21</w>
<w aa="Ncmsi" ana="Ncmsi">век</w>

<w aa="Mcfsi;Mcmsi;Mcnsi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mofsi">1991</w>
<w aa="Ncfsi" ana="Ncfsi">година</w>
```

- **Integers ending in 2** (2, 22, 32, ... without numerals ending in 12).

  In these cases the grammatical features are these of the cardinal number two or three etc. (два (Mcmpi), две (Mcfpi), две (Mcnpi), две (Mc-pi)) and the ordinal number second (втори (Momsi), втора (Mofsi), второ (Monsi), втори (Mo-pi)). Here are some examples:

```
<w aa="Mcfpi;Mcmpi;Mcnpi;Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mcmpi">2</w>
<w aa="Ncmsh;Ncmt" ana="Ncmt">грама</w>

<w aa="Mcfpi;Mcmpi;Mcnpi;Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mcfpi">2</w>
<w aa="Ncfpi" ana="Ncfpi">чаши</w>

<w aa="Mcfpi;Mcmpi;Mcnpi;Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mcnpi">2</w>
<w aa="Ncnpi" ana="Ncnpi">кафета</w>

<w aa="Vpitf-o2s;Vpitf-o3s;Vpitf-r3s" ana="Vpitf-r3s">завършва</w>
<w aa="R;Te" ana="R">на</w>
<w aa="Mcfpi;Mcmpi;Mcnpi;Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mc-pi">2</w>

<w aa="Mcfpi;Mcmpi;Mcnpi;Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Momsi">22</w>
<w aa="Ncmsi" ana="Ncmsi">век</w>

<w aa="Mcfpi;Mcmpi;Mcnpi;Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mofsi">1992</w>
<w aa="Ncfsi" ana="Ncfsi">година</w>
```

- **Other integers** (-5, 3, 10, 11, 12, 23, 33).

  In these cases the grammatical features are these of the cardinal number three (три (`Mc-pi`)) and the ordinal number third (трети (`Momsi`), трета (`Mofsi`), трето (`Monsi`), трети (`Mo-pi`)). Here are some examples:

```
<w aa="Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mc-pi">12</w>
<w aa="Ncmsh;Ncmt" ana="Ncmt">грама</w>

<w aa="Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mc-pi">12</w>
<w aa="Ncfpi" ana="Ncfpi">чаши</w>

<w aa="Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mc-pi">12</w>
<w aa="Ncnpi" ana="Ncnpi">кафета</w>

<w aa="Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Momsi">12</w>
<w aa="Ncmsi" ana="Ncmsi">век</w>

<w aa="Mc-pi;Mo-pi;Mofsi;Momsi;Monsi" ana="Mofsi">1993</w>
<w aa="Ncfsi" ana="Ncfsi">година</w>
```

- **Decimal fractions** (-5.4, 3.14159, 0.0010, 1,133).

  In these cases the grammatical features are these of the cardinal numbers different from one and two: `Mc-pi`.

```
<w aa="Mc-pi" ana="Mc-pi">1.72</w>
<w aa="Ncmsh;Ncmt" ana="Ncmt">милиона</w>
<w aa="Ncmsh;Ncmt" ana="Ncmt">лева</w>
```

- **Fractions with slash** (1/14, 2/15, 3/16, 25/8).

  In these cases the grammatical features depend on the figure before the slash. If the figure before slash ends in 1 (but not in 11) then the grammatical features are represented by the tag: `Mcfsi`. If the figure before slash ends in 2 (but not in 12) then the grammatical features are represented by the tag: `Mcfpi`. In all other cases the grammatical features are represented by the tag: `Mc-pi`.

```
<w aa="Mcfsi" ana="Mcfsi">1/2</w>

<w aa="Mcfpi" ana="Mcfpi">2/15</w>

<w aa="Mc-pi" ana="Mc-pi">23/11</w>
```

- **Roman figures** (I, V, IV, XXI, XXXVIII, XII).

  Roman figures stand only for ordinals. Thus, the grammatical features are `Momsi`, `Mofsi`, `Monsi`, `Mo-pi`. Here are some examples:

```
<w aa="Mo-pi;Mofsi;Momsi;Monsi" ana="Momsi">XXI</w>
<w aa="Ncmsi" ana="Ncmsi">век</w>

<w aa="Mo-pi;Mofsi;Momsi;Monsi" ana="Mofsi">VII</w>
<w aa="Ncfsi" ana="Ncfsi">гимназия</w>

<w aa="Mo-pi;Mofsi;Momsi;Monsi" ana="Monsi">XXXVIII</w>
<w aa="Ncnsi" ana="Ncnsi">събрание</w>

<w aa="Mo-pi;Mofsi;Momsi;Monsi" ana="Mo-pi">XII</w>
<w aa="Ncfpi" ana="Ncfpi">игри</w>
```

2.6 Punctuation

The punctuation is annotated with the element `<pt>`:

```
<pt type="Type">punctuation mark</pt>
```

The attribute `type` is applied to the classification of some uses of quotation marks. It determines whether the quotation mark is an opening one or a closing one. This fact is stated only when the use of the quotation mark can be determined automatically.

Here is the definition of the element and its attributes from the BulTreeBank DTD:

```
<!ELEMENT pt (#PCDATA) >
<!ATTLIST pt
    type CDATA #IMPLIED
>
```

3 Morpho-Syntactic Ambiguities

In this section we present some of the most frequent ambiguities at the morpho-syntactic level in Bulgarian texts. For some of them we present rules, or guidelines how to distinguish the right features in the given context of usage. In order to facilitate the task of morpho-syntactic disambiguation we follow two requirements:

- Re-modelling the traditional part-of-speech groups in order to separate the morphological categories from their usages, and
- Definition of preference rules in certain cases where more than one decisions are possible.

Concerning the first requirement, we adhered to the following general steps:

- introducing hybrid part-of-speech tags (auxiliary particle, verbal particle, adverbial numeral, adjectival name, dative-possessive clitic);
- underspecification of one of the possibilities (excluding the particle role while maintaining the conjunction one; focusing on parenthetical function in certain positions of some conjunctions like обаче (obache, however));
- making some underspecified features of the same lexeme explicit (transitivity is always separated from intransitivity as well as perfective from imperfective).

Here is the list of the main decisions:

- да ('da', to/yes) is an auxiliary and affirmative particle: `(Tx;Ta)`.
- ще ('shte', will), нека ('neka', let), ето ('eto', here is/are) in certain contexts are auxiliary particles: `(Tx)`. Note that the past form is a verb as well as the homonymic form with meaning 'want'.
- макар ('makar', although) is a preposition and a particle: `(R;Te)`.
- the quantity words много ('mnogo'), малко ('malko'), повече ('poveche') and their synonyms and derivational variants are treated as adverbial numerals: `Md--i` or `Md--d`. Some of them also have other tags.
- the words of enumerating like 'pyrvo', 'vtoro' etc. are treated as adverbs of time and ordinals `(Dt;Monsi)`.
- the tags for verbs, which are treated as both perfective and imperfective, are separated, and thus dependant on the context.
- the tags for verbs, which are treated both transitive and intransitive, are separated and thus dependant on the context.

2.1. When there is an ambiguity between a dative and a possessive reading in verb adjacent position, we select the hybrid dative-possessive option. (Toj mi vze paltoto).

When disambiguating between parts-of-speech or wordforms, the following preference rules are applied:

- When there is an ambiguity between a parenthetical (adverbs of modal nature) and conjunction reading, choose the parenthetical if conjunctive is suppressed: from `Dd;Cc` select `Dd`. Example: Той обаче не се появи. ('Toj obache ne se pojavi'), here обаче ('obache') is annotated as adverb. The preference rule says: When the word is in a non-initial sentence position, then choose the adverb tag. Otherwise, choose the conjunction one.
- When there is an ambiguity between a conjunction and particle reading, choose the conjunction, if the other is unclear or suppressed: from `C*;Te` select `C*` (C* stands for any kind of conjunction). Example: А той ще дойде ли? ('A toj shte dojde li?'), here А ('A') is annotated as a conjunction.
- When there is an ambiguity between an adverb of time and noun reading, then choose the adverb, if no inflection is shown: `Dt;Ncmsi` select `Dt`. Example: Ще дойда утре следобед. ('Shte dojda utre sledobed'), here следобед ('sledobed') is annotated as an adverb.

Several frequent types of ambiguities are listed and discussed here. These types are corpus-driven and differ from the frequent types one can find in the lexicon. They show the degrees of manageability of the morpho-syntactic ambiguity problem. For some types, strict rules can be proposed due to the existence of clear syntactic and positional indicators (1, 2, 4 etc.); for others some rules can also be proposed, but human inspection is always expected (3, 5, 9, 10, 11 etc.), i.e. some of the cases remain unresolved (or resolved wrongly) unless a human expert intervenes.

1. Ambiguities between a preposition and a particle: `(R;Te)`.

   Макар ('makar', although) is considered a preposition which takes clausal complements. When there is no clausal complement, it is a particle. (Макар `[R]` че беше тук, той не знаеше нищо. vs. Да ти се не види макар `[Te]`.)

2. Ambiguities between the auxiliary particle да ('da', to) and the affirmative particle да: ('da', yes) `(Tx;Ta)`.

   The affirmative particle typically appears in sentence initial and final position in contrast to the auxiliary particle. The possible places of ambiguity are the sentence initial and the central position, but, as a rule, the affirmative particle is always separated by comma or other final or non-final punctuation (exclamative mark, full stop, question mark, dash, comma etc.). (Да `[Ta]`, той беше дошъл. Vs. Да `[Tx]` дойдеш ли искаш?).

3. Ambiguities between the third person, singular auxiliary е ('e', is), the emphasis particle е and the interjection е: `(I;Te;Vxitf-r3s)`.

   Usually the emphasis particle comes in initial position and is separated by punctuation mark with the exception of exclamation mark: (Е `[Te]`, кой ще идва?).

   The interjection is recognized by the following exclamation mark: (Е `[I]`! Не очаквах това.).

   The verb can be only followed by comma in sentence internal position: (Той е `[Vxitf-r3s]`, който ще ме утеши.)

4. Ambiguities between the singular, masculine noun with a short article and the count plural masculine noun: `(Ncmsh;Ncmt)`.

   The singular, masculine noun with a short article is chosen after a preposition or in stand-alone position: (Той беше ранен в крака `[Ncmsh]`.; Видях крака `[Ncmsh]` му.)

   The count plural masculine noun is chosen after a numeral or a quantity adverbial pronouns like няколко ('njakolko', several), колко ('kolko', how many): (Столът има четири крака `[Ncmt]`.).

   When the noun is a head of noun phrase the form depends on the modifiers of the head. If one of the modifiers is a numeral or a quantity adverbial pronouns then the noun is count plural masculine. Example: (Там имаше няколко много високи стола`[Ncmt]`.).

   This rule can be determined also locally: if there are several adjective modifiers in front of the head noun then the noun has to agree in number with the last adjective. If this adjective is plural then the noun is count plural masculine. Note that the combination of a singular adjective and a singular, masculine noun with a short article is not possible in Bulgarian. The article (if presented) has to be attached to the first element of the noun phrase.

   Remark: An idiosyncratic form is the lexeme ден ('den', day). Very often the plural form дни ('dni', days)is used in the position of the count form дена ('dena', days) and competes it. In these cases we treat the plural form as a count one.

5. Ambiguities between 2nd person and 3rd person aorist and 3rd person present in imperfective, transitive verbs: `(Vpitf-o2s;Vpitf-o3s;Vpitf-r3s)`.

   After the auxiliary particles да ('da', to), ще ('shte', will), нека ('neka', let) 3rd person present always comes.

   Other markers for choosing present tense are the adverbs of frequency (понякога ('ponjakoga', sometimes), винаги ('vinagi', always), често ('chesto', often), etc.).

   For resolving 2nd or 3rd person within aorist, broader context sometimes is needed for the correct anchoring.

6. Ambiguities between neuter, singular adjectives and adverbs of manner: `(Ansi;Dm)`.

   The adverbs of manner cannot modify nouns. But the problem comes in predicative uses.

When there is agreement with the subject, the adjective is chosen. In some cases the ambiguity is grammatically irresolvable: (Детето влезе засмяно.). In this case we take into account the semantic criterion and choose only one of the possibilities.

7. Ambiguities between the subordinate, coordinate conjunctions че ('che', that) and the particle че: `(Cc;Cs;Te)`.

   'Che' is a conjunction always after a comma. Its coordinative usage is very rare and archaic: (Тя излезе, че (=и) `[Cc]` си китка закичи.).

   When in sentence initial position, more attention is needed. It can be either a conjunction (Тя се боеше. Че `[Cs]` ще падне.), or a particle (Че (=та) `[Te]` аз ли не знам?).

8. Ambiguities between 2nd person and 3rd person aorist, 3rd person present and imperative in perfective, transitive verbs: `(Vpptf-o2s;Vpptf-o3s;Vpptf-r3s;Vpptz-2s)`.

   The same things hold from point 5 (above).

   Additionally:

   - Sometimes broader context is needed to decide between present and aorist: (In 'Дойде, запали цигара и после тръгне...' the last verb prompts that the first two verbs are in present tense, because its present and aorist form are different.).
   - Imperatives are usually in exclamative sentences: (Уплаши го!).

9. Ambiguities between the noun, the possessive, dative pronoun, dative-possessive pronoun and the auxiliary 2 person singular verb си 'si': `(Ncnsi;Ppxtd;Ppxts;Psxto;Vxitf-r2s)`.

   The noun is domain-specific and therefore - rare. It is likely to occur in inverted commas etc., or following the noun 'note'.

   The possessive pronoun usually comes after a definite noun. The dative pronoun has usually a pre- or post verbal position.

   But in some cases it is difficult to rely only on the listed above tests: In constructions where the clitic is verb adjacent, but at the same time is connection with the nominal verb argument (Вземи си палтото) the dative-possessive tag is chosen.

10. Ambiguities between the preposition and the particle: 'по' `(R;Te)`.

    The preposition cannot be followed by non-nominals (note that under nominals we have in mind also clauses that can be nominalized) (Ти по `[Te]` вървиш от него.).

    Therefore, the difficulty comes with the nouns: (Той е по `[Te]` мъж от теб vs. Тя е по `[R]` мъж Стоянова).

11. Ambiguities between the masculine, singular, indefinite adjectives ending in -ски (-ski), their plural, indefinite form and adverbial function: `(A-pi;Amsi;Dm)`.

    The most reliable test is the number feature of the modified noun (мъжки `[Amsi]` дух `[Ncmsi]`, мъжки `[A-pi]` времена `[Ncnpi]`).

    In adverbial function these words cannot directly modify nouns.

12. Ambiguities between the coordinating conjunction, the particle and the interjection 'а': `(Cp;I;Te)`.

    The interjection is used in exclamative sentences or is delimited by punctuation (А `[I]`! Той е тук.; А `[I]`, накъде си тръгнал?).

    The distinction between the conjunction and particle is not always trivial:

    - the conjunction comes after a comma;
    - the particle comes in sentence-final position;
    - in sentence initial position it is difficult to decide. (see point 7 above)

For that reason we assume that the discourse conjunction role is more general and subsumes the role of the particle. Therefore, we choose more often conjunction.

13. Ambiguities between the neuter, singular adjective and the modal or manner adverb: `(Ansi;Dd;Dm)`.

    These adverbs do not modify nouns and do not show gender and number agreement. In predicative contexts we treat the non-verbal predicatives as adjectives, unless the form for the adverb is non-homonymic with the adjective: Вероятно `[Dd]` е да го видиш. vs. Това е вероятно `[Ansi]`. But: Добре е да дойдеш.

14. Ambiguities between 'един' (edin, 'one') as indefinite pronoun and cardinal numeral: `(Mcmsi; Pfe-os-mi)`.

    - When the form is in plural, there is no ambiguity, because it can be a form only of the indefinite pronoun.
    - When the form is in front of abstract or mass noun, choose the Indefinite pronoun.
    - When the context does not indicate the numerical meaning, choose the indefinite pronoun tag.

    The numerical meaning is present for example in the following contexts:

    ```
    Дай ми само една книга.
    Daj mi samo edna kniga.
    Give me only one book.

    Дай ми една книга, а не две.
    Daj mi edna kniga, a ne dve.
    Give me one book, but not two.

    Едната от книгите беше негова.
    Ednata ot knigite beshe negova.
    One of the books was his.

    Той беше един на майка.
    Toj beshe edin na majka.
    He was one to mother.
    He was the only child.
    ```

15. Ambiguities between adjectives, participles and nouns: `(Ansi;Ncnsi)`.

    The substantivized adjectives/participles/pronouns are marked as nouns (for ex. the moral and esthetic categories - зло ('zlo', evil), добро ('dobro', good), красиво ('krasivo', beauty) or common expressions like: болните ('bolnite', the sick), учащите ('uchashtite', the studying), домашното ('domashnoto', the homework)).

    Note that the ambiguity between neuter adjectives and adverbs is discussed above.

    More attention should be paid to the elliptical cases:

    ```
    Аз срещнах добро куче, а тя - зло [куче].
    Az sreshtnah dobro kuche, a tia - zlo [kuche].
    I met a good dog, and she - bad [dog].
    ```

16. Ambiguities between masculine and neuter in the homonymic accusative and dative forms of the short personal pronouns: `(Ppetas3m;Ppetas3n)`, `(Ppelas3m;Ppelas3n)`, `(Ppetds3m;Ppetds3n; Ppetss3m;Ppetss3n;Psot--3--m;Psot--3--n;Te)`, `(Ppelds3m;Ppelds3n)`.

It is due to insufficient context (a problem of disambiguating isolated sentences!). Examples:

(a) Видях го

There are two possible analyses where the accusative clitic 'go', depending on its referent, is either masculine or neuter:

```
Saw-I him.
I saw him.
```

or

```
Saw-I it
I saw it
```

In this case we take an opportunistic approach, namely: If the context is not sufficient, just choose the masculine option.

17. Ambiguities between adjective, adverb and possibly noun in phrases with a postmodified negative, collective or indefinite neuter pronoun: `(Ansi;Dm;Ncnsi)`.

    In such cases adjectives are chosen: нещо хубаво ('neshto hubavo', something good), нищо хубаво ('nishto hubavo', nothing good), всичко хубаво ('vsichko hubavo', all the best) → хубаво ('hubavo', good) is an adjective.

18. Ambiguities between dative clitics and particles, derived from them: `(Psot-#;Te)`.

    When the clitics have only stylistic role (for example, dativus eticus): 'onija mi ti', they are treated as particles.

4 Discussion on Some Hard Nuts

4.1 The Verbs

When the features like perfectivity/imperfectivity and transitivity/intransitivity are always separated and dependant on the context, then they need a proper treatment. Here we suggest some strategies:

1. How to distinguish between perfective and imperfective in non-homonymous cases:

   If the form can be used in present actual tense, then it is imperfective, i.e. in the context: 'At the moment I am doing something ...'. Otherwise it is perfective.

   Example:

   В момента КАЗВАМ ...(At the moment I am saying ...) → imperfective

   * В момента КАЖА ... (At the moment I say ...) → perfective

   Remark: One should be capable of deriving the correct form in present tense of the verb and not to confuse it with the counterpart verb. For example, given the form: КАЗАХ (said-I) the verb in present tense is КАЖА, but not: КАЗВАМ.

2. How to distinguish between perfective and imperfective in homonymous cases (basically in 3rd conjugation verbs):

   This decision is rather context-bound. Sometimes both aspects are acceptable. One of the tests is to substitute this verb with a non-3rd-conjugation verb and to apply the criterion, described above.

   Another clue is: when there are adverbs of frequency, it is always imperfective verb.

   Example:

   Вчера го арестуваха (Yesterday they arrested him). → perfective

   Вчера дълго го арестуваха (Yesterday they were arresting him for a long time) → imperfective

4.2 The adverbs

Some problems arise within two sub-types of adverbs, one of which is the modal adverb usage. The general rule says: if the literal meaning is expressed, the non-modal meaning is triggered. Below we present the most frequent usage pairs:

1. наистина ('naistina', indeed), `(Dd; Dm)`

   ```
   Той наистина дойде
   Toj naistina dojde (Dm)

   Той, наистина, дойде
   Toj, naistina, dojde (Dd)
   ```

2. точно ('tochno', punctually or namely), особено ('osobeno', strangely or namely), `(Dd;Dm)`

3. вече ('veche', already), `(Dd;Dm)`

   It is always `Dt`. Only the superlative form is `Dd`.

5 Conclusion

Morphosyntactic disambiguation for Bulgarian is viewed as a complex process, which includes several procedures and approaches: re-classification of certain parts-of-speech, preference rules for difficult cases as well as some opportunistic techniques like the choice of only one possibility in ambiguous contexts.

References

[Osenova and Simov 2004] Petya Osenova and Kiril Simov. 2004. *BulTreeBank Style Book.* BulTreeBank Technical Report BTB-TR05.

[Simov et. al. 2004] Kiril Simov, Petya Osenova, Milena Slavcheva. 2004. *BulTreeBank Morphosyntactic Tagset. BTB-TS version 2.0.* BulTreeBank Technical Report BTB-TR03.