Comparing Bulgarian and Slovak Multext-East morphology tagset

Ludmila Dimitrova[a)], Radovan Garabík[b)], Daniela Majchráková[b)]
a) Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
1113 Sofia, Bulgaria
ludmila@cc.bas.bg
b) Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
813 64 Bratislava, Slovakia
korpus@korpus.juls.savba.sk, http://korpus.juls.savba.sk

Abstract

We analyse the differences between the Bulgarian and Slovak languages Multext-East morphology specification (MTE, 2004). The differences can be caused either by inherent language dissimilarities, different ways of analysing morphology categories or just by different use of MTE design guideline. We describe all the parts of speech in detail with emphasis on analysing the tagset differences.

Introduction

The EC project MULTEXT *Multilingual Tools and Corpora* produced linguistics resources and a freely available set of tools that are extensible, coherent and language-independent, for seven Western European languages: English, French, Spanish, Italian, German, Dutch, and Swedish (Ide, Veronis, 1994). The EC INCO-Copernicus project MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages* is a continuation of the MULTEXT project. MULTEXT-East (MTE for short; Dimitrova et al., 1998) used methodologies and results of MULTEXT. MTE developed significant language resources for six Central and Eastern European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, as well as English. Three of these languages (Bulgarian, Czech, and Slovene) belong to the Slavic language group. The results of the two projects MULTEXT and MTE are:

- tools, corpora, and linguistic resources for thirteen western and eastern European languages, with extensions to regional languages (Catalan, Occitan) and non-European languages (Bambara, Kikongo, Swahili);
- experience of developing standards and specifications for encoding of linguistic corpora;
- experience of using the same program tools for the processing of linguistic corpora.

These results show how important the development of common, harmonised and unified resources for different European languages and the language independence of the tools employed are.

The MTE electronic linguistics resources include a multilingual corpus and datasets of language-specific resources. The language-specific resources that the MTE project developed are: morphosyntactic specifications, language-specific data, and lexica.

Bulgarian morphosyntactic specifications have been made in the frame of the MTE project, but they are based on a semantic part-of-speech classification of the traditional Bulgarian grammar.

Slovak language morphology specification compatible with the MTE tagset has been developed as a projection of the Slovak morphology tagset used at the Ľ. Štúr Institute of Linguistics (Garabík, 2006), which (pragmatically) influences some parts of the specification design.

The aim of this article is to compare the differences between Slovak and Bulgarian MTE specification. Specifically, our goal is *not* to compile a list of grammar differences between the languages – we only gloss over them as far as they influence the morphosyntactic tagsets used.

The tagset differences describer can be separated into three different categories:

1. Differences due to inherent differences between the languages. For example, Bulgarian has lost (up to few exceptions) the Proto-Slavic case system, while Slovak keeps it almost fully – subsequently, the Case attribute is present only sporadically in the Bulgarian tagset, while the Slovak Case category is ubiquitous. We include also the differences resulting from orthography tradition here, since we are primarily dealing with the written language, where the orthography forms an inseparable part of language analysis.
2. Differences due to different way of analysing the morphology, either as described by traditional grammars, or by different design decisions in our tagsets. Most notably, Slovak tagset strives to cover the morphology at the lowers possible level and assumes thorough tokenization into the smallest possible units – there are no multi word tokens in the Slovak tagset (each part of such an expression will be assigned its own tag), while in Bulgarian multiword expressions are common (e.g. Bulgarian expression ***дявол да го вземе*** will be classified as interjection, while the Slovak ***čert ho ber*** will be analysed as three independent words, noun+pronoun+verb – the two expressions are otherwise identical in both languages).
3. Different way of putting grammar information into the Multext East tagset. Since the Bulgarian and Slovak tagsets were created independently, using only the MTE guidelines as a common references, there are some features that have no base neither in the primary grammar differences, nor in traditional descriptions, but rather reflect the ambiguity of categorization of grammar features in the scope of MTE. The Slovak MTE tagset is secondary to a morphosyntactic tagset developed to analyse Slovak language in the Slovak National Corpus (Garabík et al., 2004) – in fact, there is also an automatic algorithm mapping the corpus tagset into the MTE one, therefore its design is in some points influenced by the primary tagset as well.

Several words on terminology used: Category is a part-of-speech, consisting of Noun(N), Verb(V), Adjective(A), Pronoun(P), Determiner(D), Article(T), Adverb(R), Adposition(S), Conjunction(C), Numeral(M), Interjection(I), Residual(X), Abbreviation(Y), Particle(Q). Each category has one or several attributes, and each attribute can have exactly one value (including special value '-', meaning 'not applicable'). Throughout the article, we write the one letter abbreviation of a specific category or value in parentheses after the full name. To differentiate the established meaning of *grammar* category from the MTE Category term, we always use the expression *grammar category* for the former.

Values (but not the whole categories or attributes) used only in one of the MTE languages are denoted as 'language specific' in the MTE specification and we mark them with the [l.s.] abbreviation following the value name.

Common differences

There are some features specific for both – Bulgarian and Slovak - languages, which occur repeatedly in several categories, and which we describe here at the beginning, to avoid unnecessary repetition.

Case attribute

Old-Bulgarian had an elaborate case system – there were three numbers for nouns, for example, and seven cases for each of these three numbers. In the process of development of Bulgarian from a synthetic/inflectional language to an analytic/flectional language, case forms were replaced with combinations of different prepositions with a common case form. Case forms then dropped out, and only some have remained in the language until current day. Bulgarian has lost most of the traditional old Slavic case system. For nouns, best preserved is the vocative form, which has survived in the proper names (mostly in given names and some other typically addressee nouns (***Иване, жено, народе*** */Ivan, woman, folks/*). In some local dialects, the genitive-accusative form is well preserved with proper male name noun forms: ***Тичай до Ивана, до Стояна*** (instead of ***до Иван, до Стоян***) */ to Ivan, to Stoyan/*, ***Кажи на Димитра*** (instead of ***на Димитър***) */to Dimităr/*.

Most case forms have been preserved, in a systematic form, as related to pronouns (Bulgarian Grammar, 1993). Some of the Bulgarian pronouns keep the difference in nominative(n), dative(d) and accusative(a) cases.

There are no cases anywhere else, and the Case attribute is marked as 'not applicable'.

Slovak keeps the complete case paradigm for nouns, adjectives, (nominal and adjectival) pronouns, participles, and numerals, with the old Slavic vocative surviving only in some fossilized forms (***pane, bože, otče*** */sir, god, father/*) and a new vocative emerging for some given names or close family relations (***Zuzi, Pali, oci, babi*** */Zuza, Paľo, dad, grandma/*).

Definiteness attribute

One of the most important grammatical characteristics of the new Bulgarian language which sets it apart from the rest of the Slavic languages is the existence of a definite article. The definite article is a morphological indicator of the grammatical category determination (definiteness). The definite article is not a particle (particles are a separate category of words – parts-of-speech, while the article is not a separate word), nor is it a simple suffix, but a meaningful compound part of the word. It is a word-forming morpheme, which is placed at the end of words in order to express definiteness, familiarity, acquaintance (Bulgarian Grammar, 1993). In Bulgarian, nouns, adjectives, numerals, and full-forms of the possessive pronouns and participles can acquire an article.

For singular masculine, there are two forms: a full article(f)[l.s.] and a short article(s)[l.s.]. The full article is used when a singular masculine form is the syntactic subject of the clause, otherwise a short one is used – a purely orthographic rule. The distinction of full vs. short is not made for feminine, neuter and plural forms, and we use just the yes(y) or no(n) to mark definiteness or respectively lack thereof. Therefore, the definiteness attribute can take overall 4 different values: indefinite(n), definitive(y), short article(s), full article(f).
Examples:

Feminine:

жена, жената */a woman, the woman/*

```
жена =    Ncfs-n
жената =  Ncfs-y
```

жени, жените */women, the women/*

```
жени =    Ncfp-n
жените =  Ncfp-y
```

Neutrum:

дете, детето */a child, the child/*

```
дете =    Ncns-n
детето =  Ncns-y
```

деца, децата */children, the children/*

```
деца =    Ncnp-n
децата =  Ncnp-y
```

Masculine:

мъж, мъжа, мъжът */a man, the man – short art., the man – full art./*

```
мъж =     Ncms-n
мъжа =    Ncms-s
мъжът =   Ncms-f
```

мъже, мъжете */men, the men/*

```
мъже =    Ncmp-n
мъжете =  Ncmp-y
```

Slovak lacks the definiteness attribute altogether.

Animate attribute

For Slovak, the Animate attribute can be thought of as a subattribute of the masculine gender, where the words in masculine split into two categories, the animate and inanimate one. The feminine and neuter do not have this grammar category[1]. The animate is mostly used for nouns related to persons and animals. Animals are animate in the singular, but in the plural they can be both animate and inanimate, depending on the level of human characteristics assigned to them (often metaphorically). There are some borderline cases, which can be thought of as animate or inanimate in the singular as well (*robot*, as a thinking being is mostly animate, but as a mechanical tool is inanimate), or the animate feature distinguishes homonyms (***kohútik*** */rooster/* is animate, but ***kohútik*** */water tap/* inanimate).

For Bulgarian there is no animate attribute at all, and it is marked as 'not applicable'.

Part of speech specific differences

Noun

The noun in Bulgarian possesses the grammatical categories gender, number, definiteness, and (traces of) case. The noun in Slovak possesses the categories gender, number, case, and (sometimes)

[1] Sometimes a different description is used, where all the non-masculine words are inanimate by default. This is however not according to the mainstream linguistic terminology and leads to some singularities, like the word ***žena*** */woman/* being inanimate.

animateness. In both Slovak and Bulgarian, the gender is invariable and independent of word-formation. Every noun possesses one of three grammatical genders – a masculine, feminine or neuter[2].
Nouns have a singular and plural form, i.e. grammatical meaning of singular number and grammatical meaning of plural number, determined by given suffix morphemes. While in Slovak Number=singular(s) and Number=plural(p) are the only allowed values for the Number attribute, in Bulgarian there is the third value, the so-called *count form*, marked by Number=count(t)[l.s.]. This special count form in -a/-я originates from the proto-Slavic dual form. The count form appears after a cardinal numeral form (for example, ***два*** */two/, **три*** */three/*, ***четири*** */four/* etc.) or after the adverbs ***колко*** */how many/*, ***толкова*** */that many/*, **няколко** */several, a few/* with masculine nouns that end with a consonant and that do not denote persons, for example: ***два града*** */two towns/*, ***три стола*** */three chairs/*, ***четири цвята*** */four colours/*, ***колко лева*** */how many levs/*, ***няколко броя*** */ a few copies, issues/*. The count form does not appear after other adverbs such as ***много*** /many/, ***малко*** */few/*, for example ***много столове*** */many chairs/* vs. ***три стола*** */three chairs/* (Bulgarian Grammar, 1993).

Slovak keeps full featured case morphology, while Bulgarian distinguishes only nominative(n) and vocative(v) – see the discussion on cases above.
In Slovak, there is the Animate attribute, which is completely absent from Bulgarian.
Animate is differentiated only for Gender=masculine(m) and only in these cases:
1. Type=proper(p)
2. Type=common(c) & Case=accusative(a)
3. Type=common(c) & Number=plural(p) & Case ∈ { nominative(n), accusative(a), vocative(v) }

This corresponds to situations where the animateness has influence on the morphology and/or syntax. Although the animateness could be easily (with only little homonymy) assigned to all the masculine nouns, we opted for the described, rather complicated schema in order to be consistent with other MTE languages.

Pavol = **Npmsn--y**
Žiar = **Npmsn--n**
pes = **Ncmsn**
psa = **Ncmsa--y**
psov = **Ncmpg** (genitive)
psov = **Ncmpa--y** (accusative animate, homonymous with the genitive)
psi = **Ncmpa--n** (accusative inanimate, different from the animate)
žena = **Ncfsn**
ženu = **Ncfsa**

Verb

Almost all verb forms and the related grammatical meanings that existed in Old-Bulgarian have been preserved in the contemporary Bulgarian language. Unlike Bulgarian, the other Slavic languages have

[2] It can be argued that some Slovak pluralia tantum do not follow this classification. However, in traditional grammars, a given word is always assigned (often arbitrarily and forcibly) its gender, to *make* the description fit.

considerably simplified their old verb systems. The most characteristic peculiarity of Bulgarian is its very well developed system for expressing the grammar category of tense – there are forms for nine distinct verb tenses. Another important feature of the Bulgarian verb system is the presence of mood (so-called *inferential* or *re-narrative* mood) for the expression of non-witnessed modality or second-hand information. Bulgarian verbs have the grammatical categories person, number, voice, type, tense and mood. According to their lexical meaning, verbs can be transitive and intransitive. All these featured add to the complexity of the MTE tagset for Bulgarian verbs.
Some examples:

чета = Vmia1s
чета = Vmip1s
пиша = Vmip1s
заминавам = Vmia1s
заминавам = Vmip1s

Both languages keep the so-called reflexive verbs. Reflexive verbs are formed from transitive verbs with the help of the personal reflexive pronoun ***sa, ce***, or from transitive and intransitive verbs with the personal reflexive pronoun ***si***, ***cu***, for example: ***obliekať – obliekať sa, обличам – обличам се*** */dress – dress oneself/*; ***myslieť – myslieť si, мисля – мисля си*** */think – think by oneself/*. Reflexive verbs are not marked in the MTE tagset, reflexivity is shown only implicitly by the reflexive pronoun presence.

Bulgarian has only main(m) and auxiliary(a) vaues for the Type attribute, but again, Bulgarian verbs could be easily categorised in different ways (e.g. the Bulgarian ***(аз) мога*** (described as Type=main(m)) corresponds almost exactly with Slovak ***(ja) môžem*** (described as Type=modal(o)).
Slovak differentiates main(m), auxilliary(a), modal(o) and copula(c). However, this description is highly arbitrary and does not follow the traditional Slovak grammar description in detail, rather it was made for compatibility with the MTE tagset.
Vform=participle(p) corresponds to Slovak L-participle, in Bulgarian called just the participle and is used to form the past tense or the conditional. In Bulgarian, it also includes past participle (***говорено***) / *spoken/*), but this duplication will be reworked in the near future.
Vform=transgressive(t)[l.s]. in Slovak corresponds to VForm=gerund(g) in Bulgarian – this is just a difference in description.
In Slovak, imperative can be also present in the 1st person plural (***hovorme***), in Bulgarian the imperative would be formed analytically ((***хайде) да говорим*** – (particle)+particle+verb).
In both Bulgarian and Slovak, the conditional is expressed roughly in the same way, by using a separate word ***би, by***, and the L-participle form (called just participle in Bulgarian). Slovak ***by*** is for the MTE purpose highly arbitrarily classified as a verb in conditional (Vform=conditional(c), the only such verb). No other grammar categories (person, gender, tense) are marked, purely for pragmatic reasons – to avoid the need of disambiguation. On the other hand, the Bulgarian ***би*** is classified as a full verb, Vform=active(a) (this is just a superficial difference in MTE tagset):

Slovak (lemma ***by***):

by = Vcc

Bulgarian (lemma ***бъда***):

би = Vaia2s
би = Vaia3s
бих = Vaia1s
биха = Vaia3p
бихме = Vaia1p
бихте = Vaia2p

Verbs in participle form in Bulgarian can be classified for definiteness, Slovak verbs have no definiteness attribute.

In Bulgarian, there is a language specific Tense=aorist(a), and imperfect(i) value for the Tense attribute.
Past "aorist" tense expresses a past action (event) carried out or completed in a given moment or during a given period and finished before the state of speaking. Past "imperfect" expresses a past action (event) which has been carried out during a certain time period in the past or continuing until the state of speaking.

Aorist and imperfect are completely absent from Slovak.

In Slovak, voice attribute is always Voice=active(a), because passive voice occurs only in participles, which are categorised as adjectives. According to the Bulgarian grammarians there is only active voice of verbs. In Bulgarian corpus, participles are classified as verbs, with Voice=passive(p) (past tense) or Voice=active(a) (present tense) types, which not only violates the accepted grammar specification but also creates some morpho-syntactical confusion. For example ***говорещи, говорели*** are both annotated as Vmp[pai]-p-a-n, yet both are not proper verbs to possess tense category, but participles (i.e. serve syntactic role as determiners). The Voice MSD should not serve as a proxy for the Tense MSD, which is the current case for the Bulgarian corpus. This issue requires a reworking of the Bulgarian MSDs for Verb forms.

In Bulgarian, verbs can be negated with a special particle ***не*** written separately in front of the verb. In Slovak, verbs are negated by a prefix ***ne***-, which forms an unseparable part of the verb, and the lemma of a negative verb remains negative – this is more a feature of an orthography than an inherent difference in the languages. The only exception is the negation of the verb ***byť*** */to be/,* which is formed by a special particle ***nie*** written separately in front of the verb – this will be analysed as a particle, followed by a (positive) verb lemmatised as ***byť***. In Slovak MTE, there is a Negative attribute, with (rather confusing) possible values Negative=no(n) for positive verbs and Negative=yes(y) for negative ones. Bulgarian does not have this attribute.

In Slovak, there is an Aspect attribute, which appeared in MTE in version 3. The Bulgarian tagset has been designed earlier and lacks the Aspect attribute, even if the aspect in Bulgarian is roughly the same as in Slovak (and other Slavic languages). The ambivalent aspect[l.s.] is present in a special class of verbs that have the same form in perfective and imperfective/progressive aspect (the difference is only semantic/syntactic, not morphological).

Adjectives

Slovak adjectives can have either qualificative or possessive Type.
Slovak adjectives have the degree attribute, while in Bulgarian degree is formed with a separate, auxiliary particles comparative ***по*** and superlative ***най***, written with a hyphen (***хубав***, ***по-хубав***, ***най-хубав***). This can be arguably considered just a matter of different orthography tradition, however, the Bulgarian description is justified by the adjective being always in the same form, regardless of the degree.
Gender, number and person are the same in Bulgarian and Slovak.
Slovak has a full case paradigm, while Bulgarian lacks cases (there is not even a separate vocative for the adjectives, and the attribute has empty value in MTE).
Bulgarian has definiteness.
Slovak has animateness, which is governed by the agreement between adjectives and nouns.

Pronouns

Classification of Bulgarian pronouns is according to their meaning – personal, possessive, reflexive, demonstrative, interrogative, relative, indefinite, negative and general. Bulgiaran has Type=relative(r) (e.g. ***който***), which in Slovak would be formed by two consequent pronouns (***ten, ktorý***).
All the other values are compatible, there are only differences between specific classification of pronouns.
There are some traces of cases for Bulgarian pronouns, nominative(n), dative(d) and accusative(a) for personal pronouns, and their use depends on their syntactic function in the sentence – for example 1 p. sing.: ***аз*** (nom.), ***мене, ме*** (acc.), ***мене, ми*** (dat.), etc.
Slovak has full featured case paradigm for personal, adjectival and some other pronouns.
Owner_Number has the same function in Bulgarian and Slovak, however it is not described in the Bulgarian MTE (the type is left empty).
Although the Owner_Gender could be described for 3rd person possessive pronoun, both for Slovak and Bulgarian, both the Slovak and Bulgarian MTE description leave this type empty.
Clitic is the same for Bulgarian and Slovak.
Referent_Type is personal, possessive, attributive and quantitative in Bulgarian, but only personal and possessive in Slovak – the rest of pronouns do not have this type set (Referent_Type=-), which is just a deficiency in the Slovak MTE description. Otherwise the types are quite compatible between Bulgarian and Slovak.
Syntactic_Type in Slovak can be nominal(n) or adjectival(a) (e.g. ***ktorý, môj***), which is absent in the Bulgarian language (there are no adjectival pronouns of this type). Slovak also has several quasi-adjectival pronouns classified as Syntactic_Type=a (e.g. ***tvoj***), equivalents of which do exist in Bulgarian as well, but due to lack of the clear distinction of adjectival paradigm it was not felt unnecessary to introduce this value in Bulgarian MTE.
Bulgarian has definiteness, but it is present only for the possessive and reflexive types of pronouns, and for some general pronouns. Examples include:

Possessive:
Мой – моя - моят /my/
Твой – твоя – твоят /your, 2 p. sing/
Негов – неговия – неговият /his/

Reflexive:
Свой – своя – своят, своя – своята, свое-своето, свои - своите /his, her, its, their own/

Adverb

Bulgarian has language specific Type=adjectival(a), for words like ***умно*** /cleverly, wisely, sensibly/, which are derived from adjectives.
Slovak does not differentiate these two kinds of adverbs, but this is just a difference in description.
Slovak adverbs have the degree attribute, while in Bulgarian degree is formed with a separate, auxiliary particles ***по*** and ***най*** (see the discussion of degree for adjectives).

Adposition

Both languages have only prepositions, no postpositions.
Type is always preposition(p).
Slovak can contract some preposition with the following pronoun (***preň*** instead of ***pre neho***). These are described as Formation=compound(c).
Bulgarian has no compound prepositions.
Slovak tags for prepositions have the case attribute, which marks the case the preposition binds with.
Some Slovak prepositions can be vocalized, i.e. a vowel is appended to the preposition, if a following word starts with certain consonants (***v→vo, k→ku, s→so, z→zo, nad→nado***). This vocalization is not marked in the MTE tagset at all.

Conjunction

Type is the same in Slovak and Bulgarian – coordinating(c) or subordinating(s).
In Slovak, the class of two-part conjunctions has not been introduced, thus we ignore the Formation attribute.
In Bulgarian, Formation can be either simple(s) or compound(c).

Numeral

Slovak has Type cardinal(c), ordinal(o), multiple(m) and special(s), Bulgarian only cardinal(c) and ordinal(o).
In both Bulgarian and Slovak, the numerals are divided into two main categories: cardinal (quantitative) and ordinal (qualitative). Cardinal numerals signify a numerical (quantitative) property of objects: ***jeden dom, dve ženy, tri knihy; един дом, две жени, три книги*** */one home*, *two women, three books/*. Ordinal (qualitative) numerals have an enumerating property, through which one can determine the consecutive position of an object in an ensemble of homogenous objects: ***prvý deň, druhý mesiac, tretia sekunda; първи ден, втори месец, трета секунда*** */first day*, *second month,*

third second/. Ordinal numerals cannot express degrees of comparison[3], but in Bulgarian they can accept an article (definiteness is the same in Bulgarian as for nouns). The two categories of numerals are distinguished not only by meaning, but also grammatical characteristics. Cardinal numerals do not have a grammatical gender (with the exception of ***jeden, jedna, jedno, dva, dve; един, една, едно, два, две***, which were adjectives in Old Slavic) and do not change in number (with the exception of ***jeden, jedni, jedny; един, едни***), as they determine a given quantity. Ordinal numerals change gender and number just like adjectives. In Slovak, both cardinal and ordinal numerals keep morphological cases, and ordinal numerals are marked for animateness.

According to composition, numerals can be simple, complex or compound. Simple are single word numerals: ***jeden, dva, desať, sto; един, две, десет, сто*** */one, two, ten, hundred/*, complex consist of several words fused together: ***jedenásť, dvanásť, päťsto; единадесет, дванадесет, петстотин*** */eleven, twelve, five hundred/*, while compound ones are formed from two or more separate words – in Bulgarian, numerals connected with the conjunction ***и***, like ***двадесет и пет, хиляда и двеста*** */twenty five, one thousand two hundred/*, in Slovak whenever the constituents are declinable (mostly ordinals bigger than 20) – ***dvadsiaty prvý, stoosemdesiaty druhý*** /21st, 182nd/ . In MTE tagset, this distinction is not described, and compound numerals are analysed as a sequence of several separate numerals (sometimes with the conjunction ***и).***

Example:

```
един =          Mcms-ln
сто =           Mc-p-ln
единадесет =    Mc-p-ln
единадесети =   Moms—ln

jeden =         Mcmsnl--1
sto =           Mcnpnl--f
jedenásť =      Mcnpnl--f
jedenásty =     Momsnl--fy
```

For cardinals, a number is singular only for the number 1 (***jeden, един***) and ratios.

Ratios in both Slovak and Bulgarian are compound – they are composed of two numerals: ***jedna štvrtina, една четвърт*** */a quarter/*, ***tri desatiny, три десети*** */three tenths/.* In Slovak, when the numerator equals „one“, it can be optionally left out. In Bulgarian, when the numerator is one “***единица”***, the numeral is formed using suffixes: ***–ин-а*** (***половина*** */one half/*), ***–тин-а*** (***третина /*** */one third/*). In Slovak, both numerator and denominator are analysed as two separate numerals, while in Bulgarian they are analysed as one token:

```
една-четвърт = Mcfs-ln
една-пета =    Mcfs-ln
```

In Slovak MTE, the Form attribute can be one of digit(d), roman(r), letter(l),

Bulgarian has an additional Form=m_form(m), used only for people, formed with suffix –(***и***)***ма***: ***двама, трима, петима*** */two(people), three(people), five(people)/* and Form=approx(a), used for

[3] Neverthless, in Slovak there exist comparative and superlative degrees formed from the numeral ***prvý*** */the first/* - ***prvší, najprvší***. In Bulgarian only the form ***най-първи*** is used in colloquial speech.

approximate numerals (***десетина*** */about a ten/*, ***стотина*** */about a hundred/*):

```
десетина =      Mc-p-an
стотина =       Mc-p-an
```

Nouns derived from cardinal numerals with the suffixes ***-ina, -ica, -(or)ka, -ojka, -ица, -(ор)ка, -ойка*** will be classified as regular nouns – ***единица*** */a one/,* ***stovka, стотица*** */a hundred/,* ***sedmica, седморка*** */a seven/,* ***osmica, осмица*** */an eight/.*

```
единица =       Ncfs-n
стотица =       Ncfs-n
десетка =       Ncfs-n

jednotka =      Ncfsn
stovka =        Ncfsn
desiatka =      Ncfsn
```

Bulgarian has no Class attribute. Slovak has possible values according to the cardinality of the number, definite1(1) for "one", definite2(2) for "two", definite34(3) for "three" or "four", definite(f) for "five or more", demonstrative(d) ***(toľko*** */that many/),* indefinite(i) (***niekoľko*** */several/),* interrogative(q) (***koľko/*** */how many/*). Definite1, definite2, definite34 and definite are separated according to syntactical structures the numerals impose on the governed nouns – definite1 requires the corresponding noun to be in nominative singular, definite2 in nominative plural, definite34 nominative plural, definite genitive plural.

Bulgarian equivalents of demonstrative, indefinite, interrogative are classified as pronouns of a respective Type (including relative), e.g. ***няколко ученика*** */a few students/* – indefinite pronoun + noun. or sometimes as adverbs.

Interjection

Bulgarian has Formation=simple(s) or Formation=compound(c). Compound are those consisting of two (or more) words: ***боже мой!***, ***има-няма***, ***къде-къде***, ***хайде де***, ***кой знае***, ***дявол да го вземе***. Note that some of them are written with a hyphen, but some with a space, and it is the task of the tokenizer to prepare the correct tokens.

In Slovak, corresponding interjections are mostly written together (***ktovie, dočerta, čerthovie***), but sometimes separately or with a hyphen (***dovidenia***, but also ***do videnia***, ***bum bác*** but also ***bum-bác***), and these are tokenized as several separate words and analysed as either several interjections or as a residual + interjection.

Residual

In Slovak, special 'adverb prepositions' ***(po, na, do***), encountered in expressions like ***po anglicky, na zeleno, do modra*** are classified as residuals. Traditional Slovak grammars do not like to consider them separate words, but rather see them to be different part-of-speech, mostly an adverb (see interjections above), with a space inside. In corresponding Bulgarian expressions (e.g. ***на български),*** the residual will be classified as **Sp** (preposition). This is however just a difference in grammar description, not an

inherent difference in the languages.

Abbreviation

In Slovak, trailing full stop is considered to be a separate token (punctuation character). In Bulgarian, the full stop is part of the abbreviation. Otherwise the descriptions in both languages are identical.

Particle

In the Bulgarian MTE tagset, particles are characterised by the Type attribute. Type attribute is one of negative, general, comparative, verbal, interrogative, modal.
Type=negative(z) is used for particles expressing negation (***не, ни, нито***)
Type=verbal(v) is used to form different type of verbal syntactical relationships, e.g. to create future tense (***ще*** *говориш*), or particles like ***се, да*** – Slovak uses very different verbal syntactical structures.
Type=interrogative(i) are particles used to form yes/no-questions or exclamations (***ли***, ***дали***, ***нали, нима, мигар***) – this type of particles is not present in Slovak at all.
Type=comparative(c) is for particles used to create comparatives or superlatives (***по, най***) – Slovak comparatives are formed through a morphology suffix, ***naj-*** is written together with superlatives. (this could be considered just a difference in orthography).
Type=modal(o) – used to express urge or order, mostly homonymous with other types of particles, for instance ***да, дано, нека, хайде***.
Type=general(g) is for all the other, non-specialised particles.
The Formation attribute can be either simple(s) (single word particles) or compound(c) (multiple word particles, e.g. ***хайде де***).
In the Slovak MTE tagset, we simplified our task enormously by resigning the classification attempts (which can be analysed *ad nauseam* to an arbitrary precision (Šimková, 2004)), and all the articles have the same simple tag **P**. The classification has no morphology effect anyway.

Concluding Remarks

Multext East morphological tagset attempts to describe the morphology of several languages using the same principles and the same set of tags. Ideally, the differences in the respective tagsets reflex inherent underlying differences in the languages. Our analysis show that at least between Bulgarian and Slovak, there are many differences due to different way of analysing morphology in traditional grammars, as well as different Multext East tags assigned to the same categories across languages. However, we have successfully analysed the differences and pointed out categories and attributes where the discrepancies occur. In any comparative analysis of the languages based on the Multext East morphology annotation, it is necessary to take these results into account, to reveal superficial differences not based on real dissimilarities of the languages' grammars in question.

The Multext East tagset is suitable for Slavic languages. We recommend MTE morphology tagset for annotation of corpora (parallel or comparable), either as a sole morphology tagset or in addition to an established one. However, special care needs to be taken when analysing morphology across languages, because the Multext East tagset differences are sometimes artificial, based on different grammar

description, not on real differences between the languages. There are also morphology and syntax categories that the Multext East tagset does not map the same way between the languages, and therefore cannot be used uncritically in cross-linguistic analysis.

References

1. Dimitrova, 1998: Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevič, V., and Tufis, D. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. Proceedings of COLING-ACL '98. Montréal, Québec, Canada, pp. 315-319.
2. Garabík et al, 2004: Garabík, R. Gianitsová, L., Horák, A., Šimková, M., Šmotlák, M.: Slovak National Corpus. In: Proceedings of the conference TSD 2004. Brno, Czech Republic: Springer-Verlag 2004.
3. Garabík, 2006: Garabík, Radovan: Slovak morphology analyzer based on Levenshtein edit operations. Proceedings of the WIKT'06 conference, p. 2–5. Bratislava, Slovakia, 2006.
4. Ide, Véronis, 1994: Ide, N., and Véronis, J.: Multext (multilingual tools and corpora). In: COLING'94, p. 90–96, Kyoto, Japan, 1994.
5. Ružička, 1966: Morfológia slovenského jazyka. Ed. J. Ružička. Bratislava: Vydavateľstvo Slovenskej akadémie vied 1966.
6. MTE, 2004: MULTEXT-East Morphosyntactic Specifications – version 3, edition 10th May 2004.
7. Šimková, 2004: Šimková, Mária: Funkcie častíc v komunikácii. In: Jazyk v komunikácii. Medzinárodný zborník venovaný Jánovi Bosákovi. Ed. S. Mislovičová, p. 168 – 176. Bratislava: Veda 2004.
8. Bulgarian Grammar, 1993: Главна редакция Д. Тилков, Ст. Стоянов, К. Попов. Граматика на съвременния български книжовен език. Том 2 / МОРФОЛОГИЯ. Издателство на БАН. София. (In Bulgarian).