JAZYKOVEDNÝ ČASOPIS, 2011, roč. 62, č. 1

SLOVAK MULTEXT-EAST MORPHOLOGY TAGSET

RADOVAN GARABÍK
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava

GARABÍK, Radovan: Slovak MULTEXT-East Morphology Tagset. Jazykovedný časopis, 2011, Vol. 62, No. 1, pp. 19 – 39. (Bratislava)

The article presents in short, concise form the The MULTEXT-East morphology tagset as specified for the Slovak language, in the form of MULTEXT-East V3 tables, following the description of other languages present in the MULTEXT-East project. The reasoning behind some of the design choices is explained; the tagset has been influenced by other MULTEXT-East languages and by the morphosyntactic tagset used in the Slovak National Corpus.

Acknowledgements

This morphosyntactic tagset has been inspired by the Czech language specification, created by Vladimír Petkevič, with the use of Slovak morphosyntactic classification developed by the Slovak National Corpus staff.

Introduction

The wide deployment of (big) language corpora is an indispensable resource for many different kinds of linguistic research and NLP (Neuro Linguistic Programming) tools development. While the ability to search in the corpus texts is valuable by itself, an intrinsic value of a corpus can be greatly increased by adding different kinds of linguistic annotation. Some of the commonly used annotations are part-of-speech (POS) tagging, lemmatisation and morphological annotation. Lemmatisation refers to finding a basic form of a (inflected) word, called *lemma* – this is especially useful in all kinds of lexical queries, since the users do not need to concern themselves with various word forms, but concentrate on the lexemes. Part of speech tagging assigns each words in the corpus a corresponding part of speech. Lemmatisation together with POS tagging is quite powerful and sufficient level of annotation for the English language corpora, because the part of speech and lemma contain almost all the information about the word form – what little morphology English has can be almost unambiguously seen from the lemma/word-form relationship.

Working with languages with rich inflectional morphology (such as Slovak) often requires more detailed information about word's grammar categories (such as

case, number, tense, degree, gender etc.). Morphological annotation then refers to the process and results of assigning each word such information. Two important issues arise: first, we should efficiently encode the information in a human readable form indexable by corpus processing tools (typically as a short text string), and second, we should analyse and describe the morphology/grammar categories in a sufficiently unambiguous way to make a possibility for an automatic annotation. Such an analysis usually conflates POS tagging and morphology annotation – in order to arrive at the morphological categories, we must also find out the word's POS, and it is more convenient to treat the POS as just one of the categories. The set of all the tags used to describe grammar information is called a *tagset*. Such analyses usually arise from traditional grammar descriptions – however, as it turns out, traditional grammars often pose serious problems for rigorous algorithmic description. The line between syntax and morphology (and sometimes semantics) is necessarily blurred, with traditional POS description often being dependent on syntactical features or even meaning of the word in question, some of the grammar categories are unnecessarily complicated and fine grained (and automatic tagging would not achieve any reasonable precision without very difficult and elaborate setup), Traditional grammars often deal only with high level literary language, omitting "less valued" style, in prescriptivism environments, some of the grammar categories are deliberately left out or frown upon, despite their common presence in the real language corpora. Therefore, arriving at the morphological tagset for a given language is a nontrivial task, and allowances must be made for features in conflict or neglecting traditional grammars. Often, some of the syntactical information will be present in such a description (e.g. dealing with POS that – usually – do not have morphology, such as particles, conjunctions, prepositions etc.), and we use the term *morphosyntactic annotation* for such an analysis.

At the Slovak National Corpus Department of the Ľ. Štúr Institute of Linguistics, the (Slovak language) texts in the corpora automatically lemmatised and morphosyntactically tagged. The tagset used is described at the department web page[1] and is not the same as the MULTEXT-East tagset described herein. There are two other tagsets for Slovak language widely used – one has been developed at the Faculty of Mathematics and Physics, Charles University (Hajič – Vidová-Hladká, 1997) and the second one is used in the Ajka morphology analyser (Sedláček – Smrž, 2001) developed at the Masaryk University in Brno. Each of these tagsets has been developed from and is functionally identical with its Czech language tagset original.

About MULTEXT-East

The EC project MULTEXT Multilingual Tools and Corpora produced linguistics resources and a freely available set of tools that are extensible, coherent and

[1] http://korpus.juls.savba.sk/morpho.html

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 (MTE) is a continuation of the MULTEXT project. MULTEXT-East (Dimitrova – Erjavec – Ide – Kaalep – Petkevič – Tufiş, 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.

The MTE electronic linguistics resources also include a multilingual parallel corpus (based on translations of Orwell's novel 1984) and several language-specific resources – comparable corpus, speech samples, lexica with morphosyntactic tags (covering at least the vocabulary present in the 1984 translation).

About the Slovak MULTEXT-East tagset

The Slovak language was not part of the original MULTEXT-East specification – the contents of this article describe the tagset that has been developed at the Ľ. Štúr Institute of Linguistics independently of the main MULTEXT-East project. At the time of writing this article, the MTE tagset and resources released are at the version number 3; however the release of version number 4 (that differs mostly in the internal organisation of the data and file formats, most notably converting the source files into an XML based representation) is imminent. This article describes Slovak MTE specification conforming to version 3; changes in version 4 are not very significant from a linguistic point of view.

The format of this article, especially the formatting of part of speech tables and notes intentionally closely follows the established conventions of the MULTEXT-East documentation. The article also assumes that the reader possesses knowledge about general MULTEXT-East design principles and/or knowledge about some other MULTEXT-East language tagset.

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, especially following the need to have an automatic procedure for converting Slovak language morpfology tagset into the MTE one, which sometimes resulted in a less detailed information in the Slovak MTE tagset, especially when compared with other languages.

In this article, we provide the complete morphosyntactic specification for the MTE tagset. We do not explain the format of the tables and the terminology in details, rather we refer the reader to the MULTEXT-East documentation (MTE 2004)

1. Noun (N)

1.1. Lexicon

P	ATT	VAL	C	Example	Slovak term
1	Type	common	c	kniha	všeobecné meno
		proper	p	Peter	vlastné meno
2	Gender	masculine	m	otec	mužský rod
		feminine	f	kniha	ženský rod
		neuter	n	slnko	stredný rod
3	Number	singular	s	kniha	jednotné číslo
		plural	p	knihy	množné číslo
4	Case	nominative	n	pán	nominatív
		genitive	g	pána	genitív
		dative	d	pánovi	datív
		accusative	a	pána	akuzatív
		vocative	v	pane	vokatív
		locative	l	pánovi	lokál
		instrumental	i	pánom	inštrumentál
5	Definiteness		-		
6	Clitic		-		
7	Animate l.s.	no	n	hrad	neživotné
	l.s.	yes	y	otec	životné
8	Owner_Number		-		
9	Owner_Person		-		
10	Owned_Number		-		

Notes

1. Slovak distinguishes masculine animate (Animate=yes above) and masculine inanimate (Animate=no) Gender. Masculine inanimate nouns always have the same form in the nominative and accusative case, whereas masculine animate nouns have predominantly the same form in the genitive and accusative case. Masculine animate nouns and masculine inanimate nouns differ in accusative singular, nominative (vocative) and accusative plural only.
2. Slovak distinguishes 6 cases, the locative case being obligatorily prepositional. We fully realise there is no separate vocative case in the Slovak language morphology. What we called a “vocative” in these tables is in fact syntactical role of noun when used in addressing someone, a role that is only sometimes realised morphologically and in most of the cases is identical with the form of the nominative case. The exceptions exist in case of

several nouns (fossilised forms of old Slavic vocative) and (considered substandard usage of) some proper names.

3. Verbal nouns are classified as nouns.
4. Adjectival nouns *(gazdiná, hostinský)* are classified as nouns. Sometimes the distinction between noun and adjective is not as clear as we want (typical example is *obchodný cestujúci*).

1.2 Combinations

PoS	Type	Gend	Numb	Case	Anim	Examples	
N	p	m	[sp]	any	y	Pavol	Pavlovia
N	p	m	[sp]	any	n	Žiar	Žiare
N	p	f	[sp]	any	-	Lenka	Lenky
N	p	n	[sp]	any	-	Branisko	Braniská
N	c	m	s	n	-	chlap	dub
N	c	m	s	g	-	chlapa	dubu/duba
N	c	m	s	d	-	chlapovi	dubu
N	c	m	s	a	y	chlapa	
N	c	m	s	a	n		dub
N	c	m	s	v	-	chlape	dub
N	c	m	s	l	-	chlapovi	dube
N	c	m	s	i	-	chlapom	dubom
N	c	m	p	n	y	roboti	
N	c	m	p	n	n		roboty
N	c	m	p	g	-	robotov	
N	c	m	p	d	-	robotom	
N	c	m	p	a	y	robotov	
N	c	m	p	a	n		roboty
N	c	m	p	v	y	roboti	
N	c	m	p	v	n		roboty
N	c	m	p	l	-	robotoch	
N	c	m	p	i	-	robotmi	
N	c	[fn]	s	n	-	žena	mesto
N	c	[fn]	s	g	-	ženy	mesta
N	c	[fn]	s	d	-	žene	mestu
N	c	[fn]	s	a	-	ženu	mesto
N	c	[fn]	s	v	-	žena	mesto
N	c	[fn]	s	l	-	žene	meste
N	c	[fn]	s	i	-	ženou	mestom
N	c	[fn]	p	n	-	ženy	mestá
N	c	[fn]	p	g	-	žien	miest
N	c	[fn]	p	d	-	ženám	mestám
N	c	[fn]	p	a	-	ženy	mestá
N	c	[fn]	p	v	-	ženy	mestá
N	c	[fn]	p	l	-	ženách	mestách
N	c	[fn]	p	i	-	ženami	mestami

Note

In the Combinations above, “any” is a variable standing for any admissible value.

2. Verb (V)

2.1 Lexicon

P	ATT	VAL	C	Example	Slovak term
1	Type	main	m	robiť	plnovýznamové sloveso
		auxiliary	a	mať	pomocné sloveso
		modal	o	musieť	modálne sloveso
		copula	c	byť	spona (kopula)
2	VForm	indicative	i	robím	oznamovací spôsob
		imperative	m	rob!	rozkazovací spôsob
		conditional	c	by	podmieňovacia častica
		infinitive	n	robiť	neurčitok
		participle	p	robil	príčastie
	l.s.	transgressive	t	robiac	prechodník
3	Tense	present	p	robím	prítomný čas
		future	f	budem, urobím	budúci čas
		past	s	robil	minulý čas
4	Person	first	1	robím	prvá osoba
		second	2	robíš	druhá osoba
		third	3	robí	tretia osoba
5	Number	singular	s	robím	jednotné číslo
		plural	p	robíme	množné číslo
6	Gender	masculine	m	robil	mužský rod
		feminine	f	robila	ženský rod
		neuter	n	robilo	stredný rod
*	**********	**********	*		
7	Voice	active	a	robil	aktívum
8	Negative	no	n	robím	kladné sloveso
		yes	y	nerobím	záporné sloveso
9	Definiteness		-		
10	Clitic		-		
11	Case		-		
12	Animate	l.s.	-		
13	Clitic_s		-		
14	Aspect	imperfective	p	čítať	nedokonavé
	!NEW in V2!	perfective	e	prečítať	dokonavé
	l.s.	ambivalent	a	absorbovať	obojvidové
15	Courtesy	!NEW in V3!	-		

Notes

1. The verb *byť* (E. “to be”) in all its functions is characterized as Type=c (i.e. the copula), which clearly is an oversimplification because the verb has more meanings (auxiliary etc.).

2. Auxiliary verbs (Type=a) include neither the verb *byť* (see above), nor the modal verbs, and are limited to *mať,* which in turn is always classified as an auxiliary verb, which is a gross oversimplification as well.
3. The "past participle" in Slovak is used for expressing compound active past Tense and is encoded as: VForm=p(articiple), Tense=s(past), Voice=a(ctive). Note that the Czech MTE specification has also Type=p(articiple) and Tense=p(ast), here, which is clearly a mistake, because there is no Type=p for verbs, and past tense has Tense=s.
4. "Past participle" is a term taken from the common MULTEXT-East terminology, in Slovak grammars and the Slovak morphosyntactic tagset this form is called an L-participle.
5. Adjectival active and passive participles, e.g. *stojaci* (E. "standing") or *urobený* (E. "made" or "done", cf. Note 3 above) are classified as (qualificative) adjectives.
6. Negative verbs are marked as Negative=y, whereas non-negative verbs are marked as Negative=n.
7. The term "transgressive" roughly corresponds to the term "verbal participle". The transgressives have present tense and do not distinguish any other categories.
8. Gender and Animate values correspond to those associated with Nouns and are necessary to account properly for agreement.
9. Gender manifests itself in past participles only.
10. Normally, verbs form the future tense periphrastically by auxiliary *byť* (E. "to be") plus infinitive of the main verb. In addition to the copula, there are, however, some verbs which form future tense non-periphrastically, i.e. synthetically (Verbs of motion). Such verbal forms are marked as Tense=f.
11. Some modal and auxiliary verbs do not form imperative and transgressive.
12. The voice value ('a' or 'p') is not specified for VForm=c(onditional), VForm=t(ransgressive) and VForm=n(infinitive), in which case Voice='-'.
13. Verbs form negative by prefix *ne-*, with the exception of the verb *byť* (E. "to be") which forms the negative in indicative by using separate particle *nie,* e.g. *nie je* (is not). Here, *je* would be marked as positive, because of it positive form, and *nie* would be marked as a particle, thus apart from the presence of the particle, there is no other indication of "negativeness".
14. Ambivalent aspect can be considered a conflation of a pair of homonymous verbs, one in perfective, one in imperfective aspect.

2.2 Combinations

PoS	Type	VFrm	Tens	Pers	Numb	Gend	Voic	Neg	Anim	Asp	Examples
V	m	n	-	-	-	-	a	[ny]	-	any	prať, nepraťť
V	a	n	-	-	-	-	a	[ny]	-	any	mať, nemať
V	o	n	-	-	-	-	a	[ny]	-	any	musieť, nemusieť
V	c	n	-	-	-	-	a	[ny]	-	any	byť, nebyť
V	c	c	-	-	-	-	-	-	-	any	by
V	c	i	f	[123]	s	-	a	n	-	any	budem, budeš, bude
V	c	i	f	[123]	s	-	a	y	-	any	nebudem, nebudeš, nebude
V	c	i	f	[123]	p	-	a	n	-	any	budeme, budete, budú
V	c	i	f	[123]	p	-	a	y	-	any	nebudeme, nebudete, nebudú
V	m	i	f	[123]	s	-	a	n	-	any	poletím, poletíš, poletí
V	m	i	f	[123]	s	-	a	y	-	any	nepoletím, nepoletíš, nepoletí
V	m	i	f	[123]	p	-	a	n	-	any	poletíme, poletíte, poletí
V	m	i	f	[123]	p	-	a	y	-	any	nepoletíme, nepoletíte, nepoletí
V	c	i	p	[123]	s	-	a	n	-	any	som, si, je
V	c	i	p	[123]	s	-	a	y	-	any	som, si, je, niet, *neni
V	c	i	p	[123]	p	-	a	n	-	any	sme, ste, sú
V	c	i	p	[123]	p	-	a	y	-	any	sme, ste, sú
V	m	i	p	[123]	s	-	a	n	-	any	triem, trieš, trie
V	m	i	p	[123]	s	-	a	y	-	any	netriem, netrieš, netrie
V	m	i	p	[123]	p	-	a	n	-	any	trieme, triete, trú
V	m	i	p	[123]	p	-	a	y	-	any	netrieme, netriete, netrú
V	a	i	p	[123]	s	-	a	n	-	any	mám, máš, má
V	a	i	p	[123]	s	-	a	y	-	any	nemám, nemáš, nemá
V	a	i	p	[123]	p	-	a	n	-	any	máme, máte, majú
V	a	i	p	[123]	p	-	a	y	-	any	nemáme, nemáte, nemajú
V	o	i	p	[123]	s	-	a	n	-	any	musím, musíš, musí
V	o	i	p	[123]	s	-	a	y	-	any	nemusím, nemusíš, nemusí
V	o	i	p	[123]	p	-	a	n	-	any	musíme, musíte, musia
V	o	i	p	[123]	p	-	a	y	-	any	nemusíme, nemusíte, nemusia
V	c	m	p	1	p	-	a	[ny]	-	any	buďme!, nebuďme!
V	c	m	p	2	[sp]	-	a	n	-	any	buď!, buďte!
V	c	m	p	2	[sp]	-	a	y	-	any	nebuď!, nebuďte!
V	m	m	p	1	p	-	a	[ny]	-	any	pracujme!, nepracujme!
V	m	m	p	2	[sp]	-	a	n	-	any	pracuj!, pracujte!
V	m	m	p	2	[sp]	-	a	y	-	any	nepracuj!, nepracujte!
V	c	p	s	-	s	[mfn]	a	n	-	any	bol, bola, bolo
V	c	p	s	-	s	[mfn]	a	y	-	any	nebol, nebola, nebolo
V	c	p	s	-	p	m	a	n	[yn]	any	boli
V	c	p	s	-	p	[mfn]	a	y	[yn]	any	neboli
V	a	p	s	-	s	[mfn]	a	n	-	any	mal, mala, malo
V	a	p	s	-	s	[mfn]	a	y	-	any	nemal, nemala, nemalo
V	a	p	s	-	p	[mfn]	a	n	[yn]	any	mali
V	a	p	s	-	p	[mfn]	a	y	[yn]	any	nemali
V	o	p	s	-	s	[mfn]	a	n	-	any	musel, musela, muselo
V	o	p	s	-	s	[mfn]	a	y	-	any	nemusel, nemusela, nemuselo
V	o	p	s	-	p	[mfn]	a	n	[yn]	any	museli
V	m	p	s	-	s	[mfn]	a	n	-	any	robil, robila, robilo
V	m	p	s	-	s	[mfn]	a	y	-	any	nerobil, nerobila, nerobilo
V	m	p	s	-	p	[mfn]	a	n	[yn]	any	robili
V	m	p	s	-	p	[mfn]	a	y	[yn]	any	nerobili
V	c	t	p	-	-	-	a	[ny]	-	any	súc, nesúc
V	m	t	p	-	-	-	a	[ny]	-	any	robiac

3. Adjective (A)

3.1 Lexicon

P	ATT	VAL	C	Example	Slovak term
1	Type	qualificative	f	dobrý	vlastnostné
		possessive	s	matkin	privlastňovacie
2	Degree	positive	p	dobrý	1. stupeň
		comparative	c	lepší	2. stupeň
		superlative	s	najlepší	3. stupeň
3	Gender	masculine	m	dobrý	mužský rod
		feminine	f	dobrá	ženský rod
		neuter	n	dobré	stredný rod
4	Number	singular	s	dobrý	jednotné číslo
		plural	p	dobrí	množné číslo
5	Case	nominative	n	dobrý	nominatív
		genitive	g	dobrého	genitív
		dative	d	dobrému	datív
		accusative	a	dobrého	akuzatív
		vocative	v	dobrý	vokatív
		locative	l	dobrom	lokál
		instrumental	i	dobrým	inštrumentál
6	Definiteness				
7	Clitic				
8	Animate	no	n	dobré	neživotné
		yes	y	dobrí	životné
9	Formation l.s.				
10	Owner_Number				
11	Owner_Person				
12	Owned_Number				

Notes

1. Two deverbative adjectival participles, i.e. past passive participle and present active participle are not distinguished. They are conflated in the ‚qualificative' value of the Type attribute (Type=f). Past active participle is for all practical purposes dead in Slovak, although the form sometimes appears.
2. Only qualificative (and passive participle) Adjectives can be specified for Degree.

3. The attributes Gender, Number, Case and Animate correspond to the same categories within the nouns. They are necessary for the proper account of agreement of adjectives with nouns.
4. Archaic short form of adjectives survives only in some words, and only in nominative (*dlžen*, *vinen*...). These are not distinguished.
5. The qualificative adjectives which have no degrees of comparison have the Degree value equal to p(ositive).
6. Negative adjectives have negative lemma and negativeness is not marked otherwise.
7. Adjectival nouns (*gazdiná*, *hostinský*) are classified as nouns. Sometimes the distinction between a noun and an adjective is not as clear as we want (*obchodný cestujúci*).

3.2 Combinations

PoS	Type	Degr	Gend	Numb	Case	Anim	Examples
A	s	-	m	s	n	-	otcov
A	s	-	m	s	g	-	otcovho
A	s	-	m	s	d	-	otcovmu
A	s	-	m	s	a	y	otcovho
A	s	-	m	s	a	n	otcov
A	s	-	m	s	v	-	otcov!
A	s	-	m	s	l	-	otcovom
A	s	-	m	s	i	-	otcovým
A	s	-	m	p	n	y	otcovi
A	s	-	m	p	n	n	otcove
A	s	-	m	p	g	-	otcových
A	s	-	m	p	d	-	otcovým
A	s	-	m	p	a	y	otcových
A	s	-	m	p	a	n	otcove
A	s	-	m	p	v	y	otcovi!
A	s	-	m	p	v	n	otcove!
A	s	-	m	p	l	-	otcových
A	s	-	m	p	i	-	otcovými
A	s	-	[fn]	s	n	-	otcova, otcovo/otcove
A	s	-	[fn]	s	g	-	otcovej, otcovho
A	s	-	[fn]	s	d	-	otcovej, otcovmu
A	s	-	[fn]	s	a	-	otcovu, otcovo/otcove
A	s	-	[fn]	s	v	-	otcova, otcovo/otcove
A	s	-	[fn]	s	l	-	otcovej, otcovom
A	s	-	[fn]	s	i	-	otcovou, otcovým
A	s	-	[fn]	p	n	-	otcove
A	s	-	[fn]	p	g	-	otcových
A	s	-	[fn]	p	d	-	otcovým
A	s	-	[fn]	p	a	-	otcove
A	s	-	[fn]	p	v	-	otcove!
A	s	-	[fn]	p	l	-	otcových

```
 A   s     -   [fn]   p    i    -  otcovými
 A   f   [pcs]  m     s    n    -  dobrý
 A   f   [pcs]  m     s    g    -  dobrého
 A   f   [pcs]  m     s    d    -  dobrému
 A   f   [pcs]  m     s    a    y  dobrého
 A   f   [pcs]  m     s    a    n  dobrý
 A   f   [pcs]  m     s    v    -  dobrý!
 A   f   [pcs]  m     s    l    -  dobrom
 A   f   [pcs]  m     s    i    -  dobrým
 A   f   [pcs]  m     p    n    y  dobrí
 A   f   [pcs]  m     p    n    n           dobré
 A   f   [pcs]  m     p    g    -  dobrých
 A   f   [pcs]  m     p    d    -  dobrým
 A   f   [pcs]  m     p    a    y  dobrých
 A   f   [pcs]  m     p    a    n           dobré
 A   f   [pcs]  m     p    v    y  dobrí
 A   f   [pcs]  m     p    v    n           dobré
 A   f   [pcs]  m     p    l    -  dobrých
 A   f   [pcs]  m     p    i    -  dobrými
 A   f   [pcs] [fn]   s    n    -  dobrá,   dobré
 A   f   [pcs] [fn]   s    g    -  dobrej,  dobrého
 A   f   [pcs] [fn]   s    d    -  dobrej,  dobrému
 A   f   [pcs] [fn]   s    a    -  dobrú,   dobré
 A   f   [pcs] [fn]   s    v    -  dobrá!,  dobré!
 A   f   [pcs] [fn]   s    l    -  dobrej,  dobrom
 A   f   [pcs] [fn]   s    i    -  dobrou,  dobrým
 A   f   [pcs] [fn]   p    n    -  dobré,   dobré
 A   f   [pcs] [fn]   p    g    -  dobrých
 A   f   [pcs] [fn]   p    d    -  dobrým
 A   f   [pcs] [fn]   p    a    -  dobré
 A   f   [pcs] [fn]   p    v    -  dobré!
 A   f   [pcs] [fn]   p    l    -  dobrých
 A   f   [pcs] [fn]   p    i    -  dobrými
*** **** ***** **** **** **** **** ====================================
```

4. Pronoun (P)

4.1 Lexicon

```
= ============== ============== = ================== ========================
P ATT            VAL            C Example            Slovak term
= ============== ============== = ================== ========================
1 Type           personal       p ja                 osobné základné
                 demonstrative  d ten                ukazovacie
                 indefinite     i niekto             neurčité
                 possessive     s jej                osobné privlastňovacie
                 interrogative  q kto                opytovacie
                 reflexive      x sa                 zvratné
                 negative       z nikto              záporné
                 general        g každý              vymedzovacie
- -------------- -------------- -------------------- ------------------------
```

```
2 Person         first          1 ja                  prvá osoba
                 second         2 ty                  druhá osoba
                 third          3 on                  tretia osoba
- -------------- -------------- --------------------- ------------------------
3 Gender         masculine      m on                  mužský rod
                 feminine       f ona                 ženský rod
                 neuter         n to                  stredný rod
- -------------- -------------- --------------------- ------------------------
4 Number         singular       s ty                  jednotné číslo
                 plural         p oni                 množné číslo
- -------------- -------------- --------------------- ------------------------
5 Case           nominative     n ja                  nominatív
                 genitive       g tebe                genitív
                 dative         d tebe, ti            datív
                 accusative     a teba, ťa            akuzatív
                 vocative       v ty                  vokatív
                 locative       l tebe                lokál
                 instrumental   i tebou               inštrumentál
- -------------- -------------- - ------------------- ------------------------
6 Owner_Number   singular       s môj                 jednotné číslo
                 plural         p ich                 množné číslo
- -------------- -------------- - ------------------- ------------------------
7 Owner_Gender                  -
******************************** ------------------- ------------------------
8 Clitic         no             n tebe                neskrátený tvar
                 yes            y ti                  skrátený tvar
- -------------- -------------- - ------------------- ------------------------
9 Referent_Type  personal       p sa, seba, si        zvratné základné
                 possessive     s svoj                zvratné privlastňovacie
- -------------- -------------- - ------------------- ------------------------
10Syntactic_Type nominal        n on                  syntaktické substantívum
                 adjectival     a ktorý               syntaktické adjektívum
- -------------- -------------- - ------------------- ------------------------
11Definiteness                  -
- -------------- -------------- - ------------------- ------------------------
12Animate        no             n ktoré               neživotné
                 yes            y ktorí, ktorého      životné
- -------------- -------------- - ------------------- ------------------------
13Clitic_s  l.s.                -
- -------------- -------------- -
14Pronoun_Form                  -
- -------------- -------------- -
15Owner_Person                  -
- -------------- -------------- -
16Owned_Number                  -
================ ================
```

Notes

1. Gender, Number, Case and Animate correspond to the same categories as specified for nouns. They are necessary for the proper account of agreement of adjectival pronouns with Nouns.
2. Type=reflexive encompasses all reflexive pronouns (*sa, sebe, si*, *svoj, seba*) as well as *sa* in its role as the obligatory particle of reflexive verbs. Personal and possessive reflexives are further distinguished via the Referent_

Type attribute. *Sa* in all its roles will be marked as the reflexive personal clitic pronoun.

3. Pronouns are distinguished between having a (syntactically) nominal and (syntactically) adjectival function. All pronominal types except the demonstrative and possessive one can be nominal, and all except for the personal one can be adjectival.
4. Referent_Type is used to distinguish personal reflexives (which include *sa* in all its functions) from the possesive reflexives (*svoj*).
5. Negative and general pronouns ("general" Pronouns concern the Pronouns like *všetci*, *každý* etc.) are important from the viewpoint of their syntactic distribution.
6. The Clitic attribute distinguishes clitical vs. nonclitical pronominal forms, e.g. *ti* vs. *tebe.*
7. Owner_Number concern the possessor's number.
8. Owner_Gender is not marked.
9. *ty* ("you") is usually marked as vocative (cf. description of vocative for the Noun category). Many other pronouns can be marked as vocative because of their syntactical position, e.g. in *môj bože* ("'my god"), *môj* is marked as vocative.
10. *ten, tá, to, on, ona, ono* (3rd person demostrative and personal pronouns) have separate lemmas for each gender.

4.2 Combinations

```
*** **de** ***** ***** **** **** **** **** ---- --- --- ----- =============
PoS Type Pers  Gend  Numb Case Ow_N Ow_G Clit Ref Syn Anim   Examples
*** **** ***** ***** **** **** **** **** ---- --- --- ----- =============
 P    x    -     -     -   a     -    -    y   p   n    -    sa
 P    x    -     -     -  [ga]   -    -    n   p   n         seba
 P    x    -     -     -   d     -    -    y   p   n    -    si
 P    x    -     -     -  [dl]   -    -    n   p   n    -    sebe
 P    x    -     -     -   i     -    -    n   p   n    -    sebou
 P    x    -   [mfn]  all any    -    -    n   s   a  [ny-] svoj
 P    p    1     -   [sp] all    -    -  [ny]  -   n    -    ja, my
 P    p    2     -   [sp] all    -    -  [ny]  -   n    -    ty, vy
 P    p    3   [mfn] [sp] all    -    -  [ny]  -   n    -    on, oni, ony, ona
 P    s    1   [mfn]  all all   [sp]  -    n   -   a  [ny-] môj, náš
 P    s    2   [mfn]  all all   [sp]  -    n   -   a  [ny-] tvoj, váš
 P    s    3   [mfn]  all all   [sp]  -    n   -   a  [ny-] jeho, ich
 P    s    3   [mfn]  all all   [sp]  -    n   -   a    -   jej, ich
 P    d    -   [mfn]  all all    -    -    n   -   a  [ny-] 1)
 P    i    -   [mn]    s  all    -    -    n   -   n    -   2)
 P    i    -   [mfn] [sp] all    -    -    n   -   a  [ny-] 3)
 P    q    -   [mn]    s  all    -    -    n   -   n    -   4)
 P    q    -   [mfn] [sp] all    -    -    n   -   a  [ny-] 5)
 P    z    -   [mn]    s  all    -    -    n   -   n    -   6)
 P    z    -   [mfn] [sp] all    -    -    n   -   a  [ny-] 7)
 P    g    -   [mfn] [sp] all    -    -    n   -   n  [ny-] 8)
*** **** ***** ***** **** **** **** **** ---- --- --- ----- =============
```

1) ten, tento, tamten, taký, onaký
2) niekto, nejaký, bohviekto, hocikto, hocičo, niečo
3) ktovieaký, hocijaký
4) kto, čo
5) aký, čí
6) nikto, nič
7) žiadny, nijaký
8) každý

5. Determiner (D)

Not applicable.

6. Article (T)

Not applicable.

7. Adverb (R)

7.1 Lexicon

P	ATT	VAL	C	Example	Slovak term
1	Type		-		
2	Degree	positive	p	málo	prvý stupeň
		comparative	c	menej	druhý stupeň
		superlative	s	najmenej	tretí stupeň
3	Clitic		-		
4	Number		-		
5	Person		-		

Note

1. Particles form a separate part of speech category (see below) as is customary in Slovak grammars.
2. The adverbs which have no degrees of comparison have the Degree value equal to p(ositive) similarly to adjectives.

7.2 Combinations

Pos	Type	Deg	Examples
R	-	p	dobre, krížom
R	-	c	lepšie
R	-	s	najlepšie

8. Adposition (S)

8.1 Lexicon

P	ATT	VAL	C	Example	Slovak term
1	Type	preposition	p	nad, nado	predložka
2	Formation	simple	s	na, v	predložka (pravá)
		compound	c	naňho, preň	aglutinovaná pred. + zám.
3	Case				
	(req.by prep.)	genitive	g	bez	genitív
		dative	d	proti	datív
		accusative	a	pre	akuzatív
		locative	l	v	lokál
		instrumental	i	s	inštrumentál
4	Clitic		-		

Notes

1. Slovak has only prepositions, no postpositions.
2. For the disambiguation of word forms belonging to declension parts of speech it seems necessary to include the information about the case which each preposition requires.
3. A preposition can be contracted with a pronoun; such a preposition has Formation=c(ompound).
4. Prepositions can be vocalized (*v→vo*) if following word starts with certain consonant class. These are not specially marked.
5. Compound prepositions (preposition+pronoun) do not have the case category. Arguably, this kind of words can be assigned a case rather unambiguously, more ambiguously we can classify them as pronouns. In traditional grammar, they are analysed as two separate words, i. e. two different parts of speech (preposition followed by a pronoun). We decided to keep it in the preposition category and without a case, to be consistent with other MTE typology.
6. Loanword *à* (often mistakenly written as *á*) is not considered to be a preposition (that binds with the nominative), but a particle instead.

8.2 Combinations

PoS	Type	Form	Case	Examples
S	p	s	g	bez, okrem
S	p	s	d	k, proti
S	p	s	a	pre
S	p	s	l	o, v, pri
S	p	s	i	s, pod
S	p	c	-	naň

9. Conjunction (C)

9.1 Lexicon

P	ATT	VAL	C	Example	Slovak term
1	Type	coordinating	c	a, ani	priraďovacia
		subordinating	s	aby, pretože	podraďovacia
2	Formation				
3	Coord_Type				
4	Sub_Type				
5	Clitic				
6	Number				
7	Person				

Note

The class of two-part Conjunctions has not been introduced, thus we ignore the Formation attribute.

9.2 Combinations

PoS	Type	Numb	Pers	Examples
C	c	-	-	a, i, aj, ani
C	s	-	-	že, pretože, či, aby, keby

10. Numeral (M)

10.1 Lexicon

P	ATT	VAL	C	Example	Slovak term
1	Type	cardinal	c	dva	číslovka základná
		ordinal	o	piaty	číslovka radová
		multiple	m	dvakrát	číslovka násobná
		special	s	dvojaký	číslovka druhová
2	Gender	masculine	m	dva	mužský rod
		feminine	f	dve	ženský rod
		neuter	n	dve	stredný rod

```
3 Number          singular       s druhý               jednotné číslo
                  plural         p druhí               množné číslo
- -------------- -------------- -------------------- ------------------------
4 Case            nominative     n dva, dve            nominatív
                  genitive       g dvoch               genitív
                  dative         d dvom                datív
                  accusative     a dvoch, dva, dve     akuzatív
                  locative       l dvoch               lokál
                  instrumental   i dvomi, dvoma        inštrumentál
- -------------- -------------- - -------------------- ------------------------
5 Form            digit          d 5                   arabské číslice
                  roman          r MCMVIII             rímske číslice
                  letter         l šesťdesiat          číslovka slovom
* ************** ************** * -------------------- ------------------------
6 Definiteness                   -
- -------------- -------------- -
7 Clitic                         -
- -------------- -------------- - -------------------- ------------------------
8 Class           definite1      1 jeden
                  definite2      2 dva
                  definite       f päť
                  demonstrative  d toľko
                  indefinite     i niekoľko
                  interrogative  q koľko
---------------- -------------- - -------------------- ------------------------
9 Animate         no             n piate               neživotné
                  yes            y piati               životné
- -------------- -------------- - -------------------- ------------------------
10Owner_Number                   -
- -------------- -------------- -
11Owner_Person                   -
- -------------- -------------- -
12Owned_Number                   -
================================
```

Notes

1. Numerals have been specified as a separate category because of their specific syntactic distribution. We have specified two syntactic classifications by means of the attributes Type and Class; they concern different syntactic distributions. For instance *niekoľko* (E. “several”) will be characterized as: Type: cardinal, Class: indefinite. Note that difference between pronouns and these classes of numerals is fuzzy and many words traditionally considered to be numerals are indeed classified as pronouns.
2. Among the definite numbers there are three subclasses (definite1, definite234, definite) which differ in their syntactic distribution and contain the following numerals: {1}, {2, 3,4}, {5,6,...}
3. Gender, Number and Case correspond to the same categories as specified for nouns. They are necessary for the proper account of agreement of numerals with nouns.
4. Cardinal numbers have the category Number assigned according to the cardinality of the noun they bind with. Notable exception is *nula* (zero), which binds with the plural, but has been assigned Number=s(ingular), because of the conflagration with homonymous noun-like form, which can be declined.

10.2 Combinations

```
*** **** ****  **** **** ---- ------ ----- ==================================
PoS Type Gend  Numb Case Form Class  Anim  Examples
*** **** ****  **** **** ---- ------ ----- ==================================
 M    -    -     -    -    d  [123f]   -   56
 M    -    -     -    -    r  [123f]   -   MVIII
 M    c  [mfn]   s   any   l     1     -   jeden, jedna, jedno, nula
 M    c  [mfn]   p   any   l     2     -   dva, dve
 M    c    -     p   any   l     3     -   tri, štyri
 M    c    n    [sp] any   l     f     -   päť, desať, päťdesiat
 M    c    n    [sp] any   l   [diq]   -   toľko, niekoľko, koľko
 M    o  [mfn] [spd] any   l    any  [ny-] prvý, druhý
 M    m    -     -    -    l    any    -   dvakrát, päťkrát, niekoľkokrát
 M    s  [mfn] [spd]  n    l    any  [ny-] dvoje, troje, pätoro
*** **** ****  **** **** ---- ------ ----- ==================================
```

Note

In the Combinations above, ‘any’ is a variable standing for any admissible value.

11. Interjection (I)

```
= ============= ============= = ================== ========================
P ATT           VAL           C Example            Slovak term
= ============= ============= = ================== ========================
1 Type                        - au, och            citoslovce
- ------------- ------------- - ------------------ ------------------------
2 Formation                   -
=================================
```

12. Residual (X)

```
= ============= ============= = ================== ========================
P ATT           VAL           C Example            Slovak term
= ============= ============= = ================== ========================
                              - sic, %, po, na     ostatné elementy
= ============= ============= =
```

Note

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 more like modifiers of a following adverb.

13. Abbreviation (Y)

P	ATT	VAL	C	Example	Slovak term
1	Syntactic_Type		-	atď	skratka
2	Gender		-		
3	Number		-		
4	Case		-		
5	Definiteness		-		

Note

As a result of our tokenizer, abbreviations do not contain the following full stop (the full stop is analysed as a punctuation character).

14. Particle (Q)

P	ATT	VAL	C	Example	Slovak term
1	Type		-	áno, nie	častica
2	Formation		-		
3	Clitic		-		

Note

Traditionally, Slavic languages grammars recognise a separate word class of particles. The particle classification can be very detailed and complicated, reflecting their syntactical and semantic roles (Šimková, 2004). For the sake of clarity, we refrained from any classification and do not introduce any particle categories.

Conclusion

Designing a usable morphosyntactic tagsed is a non-trivial task, and there are many ways how to analyse the grammar categories in a way suitable for formal category assignment. Indeed, often there are several competing morphosyntactic tagsets for a given language, differing not at their surface forms, but deep in the underlying analysis and mapping of grammar features. Tagsets across different languages differ even more, which complicates cross linguistic research or analysis. Given that most

of the Slavic language have their MTE specifications, MTE is probably the closest thing to a unified tagset for Slavic languages. Presented specification adds the Slovak language to the increasing set of languages that can benefit form existing MULTEXT-East tagset. The specification follows Multext-East V3; Slovak language specification has been officially included in MULTEXT-East V4, but from a linguistic point of view, the specifications are the same in V3 and V4; changes in V4 concern mostly internal data format (which is now in XML).

References

DIMITROVA, Ludmila – ERJAVEC, Tomaz – IDE, Nancy – KAALEP, Heiki Jaan – PETKEVIČ, Vladimir – TUFIŞ, Dan: Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In: Proceedings of COLING-ACL'98. Montréal – Québec 1998. pp. 315 – 319.

GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária – ŠMOTLÁK, Martin: Slovak National Corpus. In: Proceedings of the conference TSD 2004. Brno: Springer-Verlag 2004.

GARABÍK, Radovan: Slovak Morphology Analyzer Based on Levenshtein Edit Operations. In: Proceedings of the WIKT'06 conference. Bratislava: 2006. pp. 2 – 5.

HAJIČ, Jan – VIDOVÁ-HLADKÁ, Barbora: Morfologické značkování korpusu českých textů stochastickou metodou. In: Slovo a slovesnost, 1997, roč. 58, č. 4, pp. 288 – 304.

IDE, Nancy – VÉRONIS, Jean: Multext (multilingual tools and corpora). In COLING'94, Kyoto 1994. pp. 90 – 96.

Morfológia slovenského jazyka. Ed. J. Ružička. Bratislava: Vydavateľstvo Slovenskej akadémie vied 1966.

MTE 2004: MULTEXT-East Morphosyntactic Specifications – version 3, edition 10th. May 2004.

SEDLÁČEK, Radek – SMRŽ, Pavel: A New Czech Morphological Analyser AJKA. In: Proceedings of TSD. Berlin: Springer Verlag 2001. pp.100 – 107.

ŠIMKOVÁ, Mária: Funkcie častíc v komunikácii. In: Jazyk v komunikácii. Medzinárodný zborník venovaný Jánovi Bosákovi. Ed. S. Mislovičová, Bratislava: Veda 2004. pp. 168 – 176.

Résumé

V rámci projektu EC MULTEXT Multilingual Tools and Corpora vznikli voľne prístupné lingvistické zdroje a nástroje pre viaceré západoeurópske jazyky: angličtinu, francúzštinu, španielčinu, taliančinu, nemčinu, holandčinu a švédčinu. EC INCO-Copernicus projekt MULTEXT-East Multilingual Text Tools and Corpora for Central and Eastern European Languages nadviazal na predchádzajúci projekt a použitím rovnakej metodológie vyvinul podobné zdroje pre šesť ďalších jazykov: bulharčinu, češtinu, estónčinu, maďarčinu, rumunčinu, slovinčinu a angličtinu. Slovenský jazyk netvoril súčasť projektu MULTEXT-East; tagset popísaný v článku vznikol nezávisle na pôde Jazykovedného ústavu Ľ. Štúra SAV v Bratislave s cieľom dosiahnuť kompatibilitu s existujúcou špecifikáciou. Článok popisuje verziu tagsetu č. 3; v súčasnosti existuje novšia verzia č. 4, ktorá sa líši iba vo formáte uložených dát, z lingvistického pohľadu nie je medzi verziou 3 a 4 podstatný rozdiel.

Formát článku úmyselne dodržuje konvencie zaužívané v dokumentácii projektu MULTEXT-East, obzvlášť vo formátovaní tabuliek a poznámok. Predpokladáme, že čitateľ je oboznámený s projektom MULTEXT-East alebo s MULTEXT-East tagsetom nejakého iného jazyka. Predkladaný tagset bol pragmaticky ovplyvnený morfosyntaktickým tagsetom slovenského jazyka používanom pri značkovaní Slovenského národného korpusu – dokonca bol úspešne vypracovaný automatický konvertor tagsetu Slovenského národného korpusu do tagsetu MULTEXT-East.

Navrhnutie použiteľného, vnútorne konzistentného morfosyntaktického tagsetu je pomerne náročná úloha, aj s ohľadom na rôzne možnosti analýzy gramatických javov v jazyku. Takáto úloha je oveľa ťažšia, ak berieme do úvahy aj iné jazyky s cieľom zachytiť v tagsetoch ich morfologickú podobnosť (alebo odlišnosť). Keďže pre takmer všetky slovanské jazyky existuje MULTEXT-East špecifikácia, používaný MULTEXT-East tagset má pravdepodobne najbližšie k (neexistujúcemu) univerzálnemu tagsetu slovanských jazykov.