Szeged Treebank CoNLL Morphological Tagset

Contents

1. Main Properties

This is not a tagset or an annotation scheme, just a format, actually the file format of the CoNLL-2009 shared task *Syntactic and Semantic Dependencies in Multiple Languages*[1]. It is based on the international morphological annotation scheme, Morphosyntactic Description (MSD)[2]. The CoNLL tagset has been generated from the MSD codes of the Szeged Treebank 2.0 and was applied in later versions of the Szeged Treebank[3]. The original MSD codes have been converted into a linearized format of attribute-value pairs. The code at position 0 is separated as the POS tag, while the other morphosyntactic attributes are in a linear order based on the MSD positions. Non-applicable attributes have `none` value. For example, the code for the Hungarian verb form *adtad* ('you gave') is: `V SubPOS=m|Mood=i|Tense=s|Per=2|Num=s|Def=y`. The annotation system is very similar to the Universal Dependencies (UD)[4] system, but with the MSD codes.

2. POS tags

POS tags of this tagset are more or less identical to the POS tags of its origin, the MSD system. Table 1 lists all possible POS tags. Subordinate classes refine the system of POS tags, which appear as the feature `SubPOS`.

Abbreviations (`Y`) do not have any features. We found a POS tag in the Szeged Corpus, namely `Z`, which also does not have any features, and which could not be identified. Several POS classes only have SubPOS features, but no others. Articles (`T`) can be divided into two classes: definite (`SubPOS=f`) and indefinite (`SubPOS=i`). Postpositions (`S`) have one feature with one possible value: `SubPOS=t`, since adpositons in Hungarian can be postpositions only. Interjections (`I`) also have only one feature with one possible value: `SubPOS=o`, which means that the interjection is a sentence-level utterance. If is not a sentence-level utterance, this feature is not specified.

3. Features

3.1. Nominals

Nominals (nouns, adjectives, numerals, and pronouns) share similar features. There are obligatory features applied for all of them: `Cas` and `Num`.

---

[1] http://aclweb.org/anthology/W09-1201
[2] http://nl.ijs.si/ME/Vault/V3/msd/msd.pdf
[3] http://rgai.inf.u-szeged.hu/index.php?lang=en&page=SzegedTreebank
[4] http://universaldependencies.org

POS	description
N	noun
A	adjective
M	numeral
P	pronoun
V	verb
R	adverb
T	article
S	postposition
C	conjunction
I	interjection
O	other
X	foreign
Y	abbreviations

Table 1: All possible POS tags.

All possible case-marking feature–value pairs are listed in Table 2. The nominative case and the singular number are unmarked in Hungarian, therefore these are the default values here. There are 24 possible cases for nominals, however not all of them applicable for all nominals. For example, the locative case can only be applied for nouns with locative meaning, while the temporal case can only be used with nominals bearing temporal meaning.

The feature `Num` have two possible values: s for the singular and p for the plural.

feature	value	description	example
Cas	2	inessive	emberben
	3	adessive	embernél
	6	multiplicative	kétszer
	9	terminative	emberig
	a	accusative	embert
	b	ablative	embertől
	c	causative	emberért
	d	dative	embernek
	e	elative	emberből
	f	essivus-formalis	emberként, emberképpen
	g	genitive	embernek
	h	delative	emberről
	i	instrumental	emberrel
	l	locative	Győrött
	m	temporal	éjfélkor
	n	nominative	ember
	p	superessive	emberen
	q	comitative (-stUl)	kuytástul
	s	sublative	emberre
	t	allative	emberhez
	u	distributive	emberenként
	w	essive	kutyául
	x	illative	emberbe
	y	translative	emberré

Table 2: Cases of nominals.

Beside the obligatory features, nominals can have features marking the number and person of the possessor or the number of the possessee. Table 3 lists these fetures and their possible values.

Nouns Nouns have two possible subclasses: common nouns (`SubPOS=c`) and proper nouns (`SubPOS=p`).

feature	value	description	example
NumP	s	singular	kutyája
	p	plural	kutyájuk
PerP	1	1st person	kutyám
	2	2nd person	kutyád
	3	3rd person	kutyája
NumPd	s	singular	kutyáé
	p	plural	kutyáéi

Table 3: Features and values marking possession.

Adjectives   Beside the common features of nominals, adjectives have three specific features. The value of the feature `Deg` marks the degree of comparison. Table 4 lists the features of adjectives.

feature	value	description	example
SubPOS	f	...	...
Cas	...	...	...
Deg	p	positive	szép
	c	comparative	szebb
	s	superlative	legszebb
Num	...	...	...
NumP	...	...	...
PerP	...	...	...
NumPd	...	...	...

Table 4: The features of adjectives with their possible values.

Numerals   Beside the common features of nominals, numerals have two extra features regarding their type and semantics. `SubPOS` feature indicates if the word is an ordinal, cardinal, fractal or distributive numeral. Table 5 lists the features of numerals and their possible values. The feature `Form` indicates if the numeral consists of letters, digits or Roman numbers.

feature	value	description	example
SubPOS	c	cardinal	három
	o	ordinal	harmadik
	f	fractal	harmad
	d	distributive	három-három
Num	...	...	...
Cas	...	...	...
Form	l	letter	három
	d	digit	3
	r	roman	III
NumP	...	...	...
PerP	...	...	...
NumPd	...	...	...

Table 5: The features of numerals with their possible values.

Pronouns   `SubPOS` feature indicates the type of the pronoun. Table 6 lists the possible values of these features.

feature	value	description	example
SubPOS	p	personal	te
	s	possessive	saját
	y	reciprocal	egymás
	x	reflexive	maga
	q	interrogative	ki
	r	relative	aki
	d	demonstrative	ez
	g	general	minden
	i	indefinite	valaki
Per	1	1st person	én
	2	2nd person	te
	3	3rd person	ő
Num	s	singular	én
	p	plural	mi
Cas	...	...	...
NumP	...	...	...
PerP	...	...	...
NumPd	...	...	...

Table 6: The features of pronouns with their possible values.

3.2. Verbs

Main verbs and auxiliaries share a common feature set, which is listed in Table 7. Infinitives do not have `Tense` and `Def` features. In the case of impersonal infinitives, `Per` and `Num` features take `none` value, while personal infinitives take person and number values.

feature	value	description	example
SubPOS	m	main verb	üt
	a	auxiliary	fog
Mood	i	indicative	üt
	m	imperative	üss
	c	conditional	ütne
	n	infinitive	ütni
Tense	p	present	üt
	s	past	ütött
Per	1	1st person	ütök
	2	2nd person	ütsz
	3	3rd person	üt
Num	s	singular	ütök
	p	plural	ütünk
Def	y	definite	ütöm
	n	indefinite	ütök

Table 7: The features of main verbs and auxiliaries with their possible values.

3.3. Adverbials

`SubPOS` feature indicates the type of the adverbial. Similarly to adjectives, adverbials can also have the `Deg` feature. Table 8 lists the possible features and values of adverbials.

Features `Per` and `Num` are always specified if it is an inflected adverbial (`SubPOS=l`). Otherwise these features take `none` value. Table 9 shows the combination matrix of adverbials.

feature	value	description	example
SubPOS	x	adverb	későn
	d	demonstrative	itt
	i	indefinite	valahogy
	g	general	bárhogy
	q	interrogative, clitic	miért, -e
	r	relative	amint
	l	inflected	mögüle
	m	modifier	sem
	p	verb particle	el
	v	gerund	futva
Deg	p	positive	későn
	c	comparative	később
	s	superlative	legkésőbb
Num	s	singular	önmagában
	p	plural	önmagukban
Per	1	1st person	magamban
	2	2nd person	magadban
	3	3rd person	magában

Table 8: The features of adverbials with their possible values.

POS	SubPOS	Deg	Num	Per
R	digqrmpv	-	-	-
R	x	[pcs]	-	-
R	l	-	[sp]	[123]

Table 9: Possible combinations of values in the case of adverbials.

3.4. Conjunctions

Table 10 shows the features of conjunctions with their possible values.

feature	value	description	example
SubPOS	c	coordinating	és
	s	subordinating	hogy
Form	s	simple	de
	c	compound	vagy...vagy
Coord	p	phrase	hogy
	w	word	de

Table 10: The features of conjunctions with their possible values.