magyarlac 2.0 MSD Morphological Tagset

Contents

1. Main Properties

MSD provides harmonised lexical specifications for ten languages, including Hungarian. The description of version 3.0 is available[1].

Morphological information is represented in attribute-value pairs, where attributes are marked by positions and values are represented by a single character. The non-applicability of a given attribute is marked by a hyphen. Position 0 encodes part-of-speech, other positions encode other morphological attributes, such as person, number, case. For example, the code for the Hungarian verb form *adtad* ('you gave') is `Vmis2s--y`. It is a main verb in indicative mode, past tense, second person singular, definite conjugation.

This tagset was used in Szeged Treebank 1.0 and 2.0, and this was the output formalism of versions 1.0 and 2.0 of magyarlanc, a toolkit for linguistic processing of Hungarian, as well. Later, a harmonized MSD-KR tagset has been developed, which is a slightly modified version of the original MSD. This tagset is used in Szeged Corpus and Treebank 2.5 and in magyarlanc 2.0. Here we refer to the latter version as MSD.

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 tags. Subordinate classes refine the system of POS-tags which is always takes the second position.

Articles (`T`) are divided into two classes indicated in the second position: definite (`f`) and indefinite (`i`). Postpositions (`S`) have one feature with one possible value: `t` in the second position, since adpositons in Hungarian can be postpositions only. The original MSD system uses two types of interjection: mood (`m`) and other (`o`), the system desribed here uses only type `o` which means that the interjection is a sentence-level utterance. If is not a sentence-level utterance, this feature is not specified. Foreign words and other residual words get only a `POS` feature, with value `X`.

3. Features

Morphological information is represented in attribute-value pairs, where attributes are marked by positions and values are expressed by a single character. The non-applicability of a given attribute is marked by a hyphen. In contrast with the original MSD system the length of the tag is not fixed, it is cut down after the last specified position.

[1] http://nl.ijs.si/ME/Vault/V3/msd/msd.pdf

POS	
N	common noun
A	adjective
M	numeral
P	pronoun
V	verb
R	adverb
T	article
S	postposition
C	coordinating conjunction
I	interjection
K	punctuation
O	other
X	foreign

Table 1: Possible values of Case feature.

There are some features that never get value in Hungarian. For example in the case of nominals and verbs there is no point in talking about gender. Due to the heritage of MSD, in where these features take fixed positions in the tag, their value is always taken by a hyphen.

3.1. Nominals

Table 2 lists all possible cases a nominal (noun, adjective, pronoun or numeral) can get.

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

Table 2: Cases of nominals.

Features regarding the person and number of possessor, and number of possessee can also occur with nominals, see Table 3.

feature	value
Owner_Number	-sp
Owner_Person	-123
Owned_Number	-sp

Table 3: Features and values marking possession.

Nouns In contrast with the MSD system common nouns and proper names are not divided, every noun gets value `n` for feature `Type`.

`POS`, `Type`, `Number` and `Case` features are always filled (and the third position is always a hyphen), consequently the minimal length of the tags of nouns is 5 characters. If the noun is a possessor, both `Owner_Number` and `Owner_Person` are filled, then the length of the tag is minimum 10 character; and if the number of the possessee appears in the inflection of the noun, it gets plus one feature reaching the maximal legth of the tag (11 characters). For all possible features and values see Table 4.

	feature	value
1	POS	N
2	Type	n
3		-
4	Number	sp
5	Case	abc...
6		-
7		-
8		-
9	Owner_Number	-sp
10	Owner_Person	-123
11	Owned_Number	-sp

Table 4: Features and their possible values of nouns.

Adjectives Beside the common nominal features adjectives have one additional feature indicating the degree of the adjective. In contrast with MSD, in which the adjective tag does not indicate if it is a participle or not, three additional values are taken in for `Type` feature for the three types of participles. See Table 5 for these features of adjectives.

Type		
f	qualificative adjective	zöld
p	present participle	mosó
s	past participle	mosott
u	future participle	mosandó
Degree		**example**
p	positive	zöld
c	comparative	zöldebb
s	superlative	legzöldebb

Table 5: Values of degree.

`POS`, `Type`, `Number`, `Degree` and `Case` features are always filled (the fourth position is always taken by a hyphen), consequently the minimal length of the tags of adjectives is 6 characters. If the adjective is a possessor, both `Owner_Number` and `Owner_Person` are filled, then the length of the tag is minimum 12 character; and if the number of the possessee appears in the inflection of the adjective, it gets plus one feature reaching the maximal legth of the tag (13 characters). For all possible features and values see Table 6.

	feature	value
1	POS	A
2	Type	fpsu
3	Degree	pcs
4		-
5	Number	sp
6	Case	abc...
7		-
8		-
9		-
10		-
11	Owner_Number	-sp
12	Owner_Person	-123
13	Owned_Number	-sp

Table 6: Features and their possible values of adjectives.

Numerals The value of feature `Type` indicates the numeral type. The original MSD code system does not use the type double for numerals, instead it uses collect number type. The value of feature `Form` indicates the calligraphical form of the numeral. See Table 7 for these features of numerals.

Type		example
c	cardinal	három
o	ordinal	harmadik
f	fractal	harmad
d	double	három-három
Form		**example**
l	letter	hat
d	digit	6
r	roman	VI

Table 7: Values of numeral type feature.

`POS`, `Type`, `Number`, `Case` and `Form` features are always filled (and the fourth position is always taken by a hyphen), consequently the minimal length of the tags of numerals is 6 characters. If the numeral is a possessor, both `Owner_Number` and `Owner_Person` are filled, then the length of the tag is minimum 12 character; and if the number of the possessee appears in the inflection of the numeral, it gets plus one feature reaching the maximal legth of the tag (13 characters). For all possible features and values see Table 8.

Pronouns The value of feature `Type` indicates the pronoun type, see Table 9. There is one additional type compared to the original MSD code system, which does not use general pronoun type.

`POS`, `Type`, `Person`, `Number` and `Case` features are always filled (and the fourth position is always taken by a hyphen), consequently the minimal length of the tags of pronouns is 6 characters. If the pronoun is a possessor, both `Owner_Number` and `Owner_Person` are filled, then the length of the tag is minimum 16 character (with 8 features getting '-'); and if the number of the possessee appears in the inflection of the pronoun, it gets plus one feature reaching the maximal legth of the tag (17 characters). For all possible features and values see Table 10.

However the code system is not hierarchical, Table 11. shows the possibe combinations of the features and their values of pronouns, because the type of the pronoun influences the possible values of some other features.

3.2. Verb

Table 12 lists the types of verbs. In contrast with the original MSD system, which uses two types of verbs (main and auxiliary), here other types are used for derived forms and their combinations.

	feature	value
1	POS	M
2	Type	cdof
3		-
4	Number	sp
5	Case	abc...
6	Form	ldr
7		-
8		-
9		-
10		-
11	Owner_Number	-sp
12	Owner_Person	-123
13	Owned_Number	-sp

Table 8: Features and their possible values of numerals.

Type		example
p	personal	te
s	possessive	övé
d	demonstrative	az
i	indefinite	valaki
g	general	bármi
q	interrogative	ki
r	relative	aki
x	reflexive	magam
y	reciprocal	egymás

Table 9: Values of pronoun type feature.

	feature	value
1	POS	P
2	Type	pdrqyxsgi
3	Person	123
4		-
5	Number	sp
6	Case	abc...
7	Owner_Number	-sp
8		-
9		-
10		-
11		-
12		-
13		-
14		-
15		-
16	Owner_Person	-123
17	Owned_Number	-sp

Table 10: Features and their possible values of pronouns.

In the case of main verbs four moods are used as the value of feature `VForm`. Indicative mood is the default and unmarked value. Table 13 lists possible values of verbal features.

`POS`, `Type` and `VForm` features are always filled, consequently the minimal length of the tags of verbs is 3

POS	Type	Pers	Num	Case	Owner_Nr	Owner_Pers	Owned_Nr
P	[px]	[123]	[sp]	[...]	[sp]	[123]	[sp]
P	y	3	s	[...]	[sp]	[123]	[sp]
P	s	[123]	[sp]	[...]	-	-	-
P	[digqr]	3	[sp]	[...]	[sp]	[123]	[sp]

Table 11: Possible combinations of values in the case of a pronoun.

Type		example
m	main verb	présel
o	potential[2]	préselhet
f	frequentative	préselget
s	causative	préseltet
1	frequentative+potential	préselgethet
2	causative+potential	préseltethet
3	causative+frequentative	préseltetget
4	causative+frequentative+potential	préseltetgethet
a	auxiliary	fog

Table 12: Values of verb type feature.

VForm		example
i	indicative	mos
m	imperative	moss
c	conditional	mosnál
n	infinitive	mosni
Tense		**example**
p	present	írok
s	past	írtam
Definiteness		**example**
y	definite	hívok
n	indefinite	hívom
2	definite, 2nd pers. object	hívlak

Table 13: Values of some verb features.

characters. If Definiteness is specified (in some types and forms of verbs) the maximal length of a verb tag is 10 characters. For all possible features and values see Table 14.

However the code system is not hierarchical, Table 15 shows the possibe combinations of the features and values of pronouns, because the type of the pronoun influences the possible values of some other features.

3.3. Adverbs

There are a lot of differences between the original MSD code system and the one described here. Table 16 lists the used types together with the information that which type is used in the original MSD as well. This solution preserves the pronominal attributes (e.g. person and number) in the adverbial use of pronouns. A type of adverbs, where a possessive suffix nominalizes the verb is used only in the original MSD system, in which the Person and Number features were compatible only with this type of adverb. In the system described here Person and Number features are usable in the inflected type of adverbs (`l`).

In contrast with the original MSD system degree feature is used in the case of adverbs, see Table 17.

The modifier '-e' question word is the only Hungarian clitic and it is marked with a `y` value of the `Clitic` feature. Some adverbs can have `Person` and `Number` values as well.

`POS` and `Type` features are always filled, consequently the minimal length of the tags of adverbs is 2 characters.

	feature	value
1	POS	V
2	Type	mofsa1234
3	VForm	imcn
4	Tense	-ps
5	Person	-123
6	Number	-sp
7		-
8		-
9		-
10	Definiteness	-yn2

Table 14: Features and their possible values of verbs.

POS	Type	VForm	Tense	Person	Number	Definiteness
V	[mofs1234]	i	[ps]	[123]	[sp]	[yn]
V	[mofs1234]	c	p	[123]	[sp]	[yn]
V	[mofs1234]	m	p	[123]	[sp]	[yn]
V	[mofs1234]	i	[ps]	1	s	2
V	[mofs1234]	[cm]	p	1	s	2
V	[mofs1234]	n	-	-	-	-
V	[mofs1234]	n	p	[123]	[sp]	-
V	a	i	p	[123]	[sp]	n

Table 15: Possible combinations of values in the case of a verb.

MSD	Type		example
	x	adverb	későn
	d	demonstrative	itt
	i	indefinite	valahogy
g	g	general	bárhogy
q	q	interrogative	miért
	r	relative	amint
	l	inflected	mögüle
m	m	modifier	sem
p	p	particle	-e
v	v	gerund	futva
o		causal	jöttömben

Table 16: Values of adverb type feature.

Degree		example
p	positive	korán
c	comparative	korábban
s	superlative	legkorábban

Table 17: Values of degree feature.

`Degree`, `Clitic`, `Number` and `Person` can grow the tag for maximum 6 characters long. For all possible features and values see Table 18.

However the code system is not hierarchical, Table 19 shows the possibe combinations of the features and their values of pronouns, because the type of the pronoun influences the possible values of some other features.

	feature	value
1	POS	R
2	Type	xdigrpvmql
3	Degree	-pcs
4	Clitic	-y
5	Number	-sp
6	Person	-123

Table 18: Features and their possible values of adverbs.

POS	Type	Degree	Clitic	Person	Number
R	[gpvmri]	-	-	-	-
R	[dx]	[pcs]	-	-	-
R	q	y	-	-	-
R	l	-	-	[sp]	[123]

Table 19: Possible combinations of values in the case of an adverb.

3.4. Conjunction

Table 20 shows the features of conjunctions with their possible values.

Type		example
c	coordinating	és
s	subordinating	hogy
Formation		example
s	simple	de
c	compound	vagy...vagy
Coord_Type		example
p	phrase (sentence)	hogy
w	word	és

Table 20: Features and their values of conjunctions.

For all possible features and values see Table 21.

	feature	value
1	POS	C
2	Type	cs
3	Formation	sc
4	Coord_Type	pw

Table 21: Features and their values of conjunctions.

3.5. Punctuation

Some punctuation marks are their own morphological code, listed together with their ASCII-code in Table 22, other punctuation marks get `K` value for POS feature.

POS	ASCII	
!	33	!
,	44	,
-	45	-
.	46	.
:	58	:
;	59	;
?	63	?
–	8211	–
others		K

Table 22: Punctuation marks that are their own morphological code.

3.6. Others

POS `O` is for words containing different punctuation marks. Feature `Type` indicates if the word is made of letters, numbers or both, and words begining or ending with a hyphen form a separate class.

Type		example
n	numbers	85-90
e	letters	index.hu
i	numbers and letters	B1-ről
h	words with hyphen	kavics-

Table 23: Possible types of words containing different punctuation marks.

The type of the word listed above determines the possible subtype categories. Table 24 lists the possible subtypes regarding the type.

SubType (numbers)		example
p	percent	7%-át
m	x or yen	800¥600
t	time	18:00-ra
f	formula	48-35
r	result	7:3-ra
d	dotted fraction	8.000
s	signed	-2
q	colon	50:50
SubType (letters)		**example**
p		A:
o	domain, slash	.org, /all
w	web address	amazon.com

Table 24: Subtypes of words containing different punctuation marks.

For all possible features and values see Table 25.

However the code system is not hierarchical, Table 26 shows the possibe combinations of the features and their values of this word class.

	feature	value
1	POS	O
2	Type	nehi
3	Subtype	-pmtfrdsowmq
4		-
5	Number	-sp
6	Case	-abc...
7		-
8		-
9		-
10	Owner_Number	-sp
11	Owner_Person	-123
12	Owned_Number	-sp

Table 25: Features and their values of words containing different punctuation marks.

POS	Type	SubType	Number	Case	Owner_Nr	Owner_Pers	Owned_Nr
O	n	p	[-sp]	[-abc...]	[-sp]	[-123]	[-sp]
O	n	[mtfrdsq]	[-sp]	[-abc...]	-	-	-
O	e	w	[-sp]	[-abc...]	[-sp]	[-123]	[-sp]
O	e	[po]	[-sp]	[-abc...]	-	-	-
O	i	-	[-sp]	[-abc...]	[-sp]	[-123]	[-sp]
O	h	-	-	-	-	-	-

Table 26: Possible combinations of values in the case of words containing different punctuation marks.