The annotation system of HunMorph

REBRUS PÉTER, KORNAI ANDRÁS, VAJDA PÉTER [1]

1 Introduction

The annotation system for Hungarian morphology was designed to satisfy at least three, sometimes contradictory, conditions. The annotation has to be

- informative: it has to reflect the morphological information of a given word-form,
- adequate: it should use linguistically adequate categories, and
- simple: easily processable by machines and humans as well.

These conditions are difficult to fulfill simultaneously. Being simple is opposed to both being adequate and informative, on the other hand the conditions mostly depend on the users' aim, whether they use the annotation system for spell-checking, stemming, syntactic analysis or statistical research.

2 Representing inflectional information as trees

The morphological description of a word has to include every inflectional feature of a given word-form. Most inflectional features play a role in syntactic analysis. Such morphosyntactic features are usually represented in an attribute-value-structure (AVS) [3]. An AVS is independent of both the the surface form of the word and the formal features of the morphosyntactic properties.

In attempting to align the above conditions we chose not to make a decision in the question of morphological segmentation. Whether we treat a morph as a whole or segment it into as many parts as the number of the morphemes it represents is a question of the chosen morphological framework. E.g. the morph '*-jaim*', though corresponds to more than one morphological property (1st person, singular possessor and plural possessed), these properties cannot be unambiguously associated with separate parts of the morph. Therefore, our annotation system does not employ the notion of segmentation in the case of suffixes. This way the annotation could be both theory neutral and modular, furthermore, it remains independent of the surface form of the word.

The morphological features of a word-form have two important properties with regard to the annotation system. The features are

- hierarchical, i.e. certain features require the presence of other features,
- asymmetrical, i.e. certain values of a feature are considered marked, while others unmarked.

[1] Documentation of LDC LCTL project

These properties are best expressed by labelled trees. The roots of the trees represent the equivalence classes of lexical entries with regard to inflection (these correspond to part-of-speech categories) and the vertices are the inflectional features. The vertices in the graph define a path with the positive values of the features. This means that the graph is capable of encoding a binary attribute-value-structure where a vertex can have a daughter only if it has positive value [2][2]. The labelled tree satisfies all three conditions. It is

- informative, as it represents morphological information in an AVS,
- adequate, as it captures morphological markedness and the hierarchical nature of inflectional information, and
- simple, as it can be automatically transformed into an AVS, furthermore, it can easily be linearized.

3 POS categories of HunMorph

The valid POS categories ar listed in Table 1. Inflectable categories are: ADJ, NOUN, NUM and VERB. The following categories cannot be inflected: ADV, DET, ART, UTT-INT, CONJ, PREV, ONO, PUNCT and PREP. For postpositions see Section 7.2.

Tag	POS category
ADJ	adjective
ADV	adverb
ART	article
CONJ	conjunction
DET	determiner
NOUN	noun
NUM	numeral
ONO	onomatopoeic
POSTP	postposition
PREP	preposition
PREV	preverb
PUNCT	punctuation
UTT-INT	utterance/interjection
VERB	verb

Table 1: POS categories of HunMorph

4 Encoding inflectional information of nouns and nominal categories

An actual feature set was designed following the above considerations for the morphological analysis of Hungarian.

[2]This is a special interpretation of markedness.

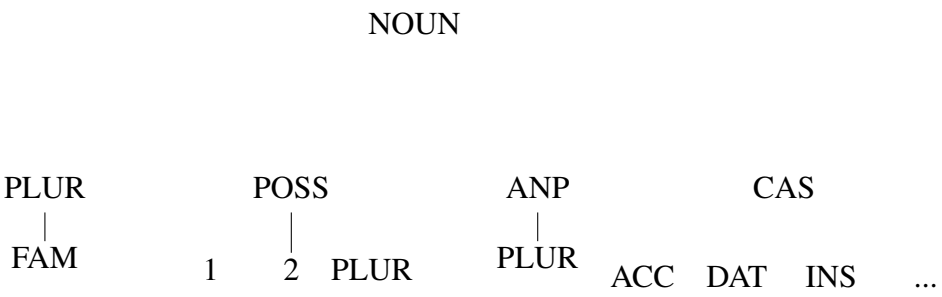

Figure 1: The signature of the graphs originating from the root node *NOUN*

In the case of a noun four binary features have to be specified. They are $\pm PLUR$ (number), $\pm POSS$ (possessor), $\pm ANP$ (possessed) and $\pm CASE$[3]. All of these can be continued as specified in Figure 1. and in Table 4. Adjectives and numerals can take the same set of inflections as nouns.

The following restrictions apply to the combination of the features:

- the $\pm CASE$ feature has to be continued by one of 16 cases,
- the $\pm PLUR$, $\pm POSS$ and $\pm ANP$ features can be contitued or can appear on their own,
- the features ± 1 and ± 2 exclude each other,
- if the $\pm PLUR$ feature of $\pm POSS$ is positive, then the $\pm FAM$ feature cannot be positive,
- if the $\pm PLUR$ and the $\pm POSS$ feature are positive simultaneously, then the $\pm FAM$ feature cannot be positive.

The morphosyntactic annotation of an inflected word-form is represented by a sub-tree of the above tree. The paths originate from the root and they encode the positive values of the attribute-value matrix. The negative values of the signature are not present in the tree. The tree is thus equivalent to an AVS encoding the inflectional properties of a word-form, however, it is free of redundancy and can be easily linearized by bracketing the nodes of the tree.

We present some examples with their full inflectional specification as an AVS and the linearization of their (sub)tree as it appears in the analysis where the outermost brackets and the + signs are omitted and the POS category is preceded by a slash and the lemma of the word-form.

kutya 'dog'

```
<NOUN<-PLUR><-POSS><-ANP><-CAS>>
kutya/NOUN
```

kutyának 'for/to the dog'

```
<NOUN<-PLUR><-POSS><-ANP><+CAS<+DAT>>>
kutya/NOUN<CAS<DAT>>
```

kutyáink 'our dogs'

```
<NOUN<+PLUR<-FAM>><+POSS<+1><-2><+PLUR>><-ANP><-CAS>>
kutya/NOUN<PLUR><POSS<1><PLUR>>
```

kutyáéi 'those things of the dog'

```
<NOUN<-PLUR><-POSS><+ANP<+PLUR>><-CAS>>
kutya/NOUN<ANP<PLUR>>
```

[3]There are two more morphosyntactic features that are in fact part of this tree. These are $\pm PERS$ and $\pm POSTP$, which are discussed in sections 7.1 and 7.2 respectively.

kutyáikéit 'those things of their dogs.ACC'
```
<NOUN<+PLUR<-FAM>><+POSS<-1><-2><+PLUR>><+ANP<+PLUR>><+CAS<+ACC>>>
kutya/NOUN<PLUR><POSS<PLUR>><ANP<PLUR>><CAS<ACC>>
```

5 Encoding inflectional information for verbs

A maximal verbal word-form has to have several properties specified. The properties are specified in Figure 2.[4] and in Table 5. The following restrictions apply to the combination of the features:

- only one of $\pm SUBJUNC$ and $\pm COND$ can be positive simultaneously,
- the feature $\pm PAST$ can only be positive if both $\pm SUBJUNC$ and $\pm COND$ are negative,
- if the feature $\pm OBJ$ is positive than its daughter feature has to positive as well,
- the feature $\pm INF$ can only combine with the feature $\pm PERSON$ $\pm PLUR$ and $\pm MODAL$.

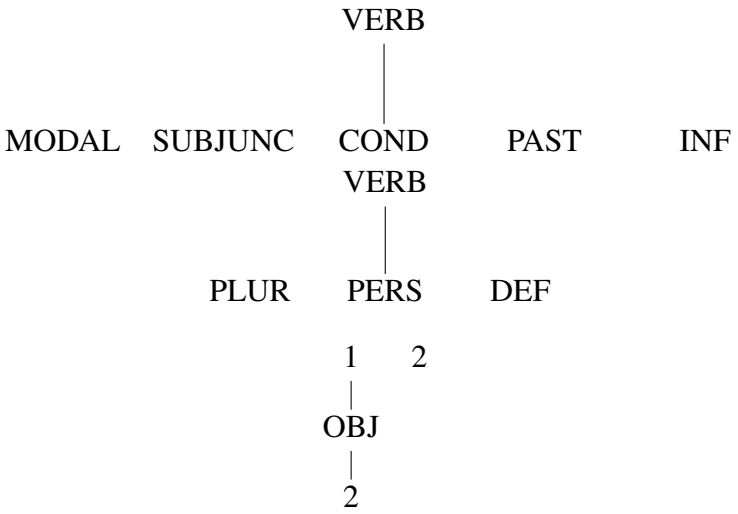

Figure 2: The signature of the graphs originating from the root node $VERB$

The annotation of verbs with inflectional suffixes is similar to that of nouns. Examples are:

lát 'he sees'
```
<VERB<-INF><-MODAL><-PAST><-COND><-SUBJ-IMP><-PERS><-PLUR><-DEF>>
lát/VERB
```
láttál 'you saw'
```
<VERB<-INF><-MODAL><+PAST><-COND><-SUBJ-IMP><+PERS<+2>><-PLUR><-DEF>>
lát/VERB<PAST><PERS<2>>
```
láthassátok 'that you may see it'
```
<VERB<-INF><+MODAL><-PAST><-COND><+SUBJ-IMP><+PERS<+2>><+PLUR><+DEF>>
lát/VERB<MODAL><SUBJUNC><PERS<2>><PLUR><DEF>
```

[4]The tree has been cut into two parts for reasons of clarity.

6 Derivation and compunding

6.1 Representing derivational information

The above tree structure is not directly suited to decribe derivation. However, a derivational suffix can be treated as a relation between two lexical entries. This way we can extend the tree structure by representing derivation as a directed edge between nodes of inflectional categories (roots of trees). Derivation can change or leave intact the POS category of a word. The POS category of the resulting word is the output category of the last derivational suffix, and the derivated word can undergo further inflectional suffixing. Inflected forms, however, cannot be subjected to derivation. Consider the following examples:

fax	`fax/NOUN`	'fax'
faxol	`fax/NOUN[ACT]/VERB`	'to send a fax'
faxolás	`fax/NOUN[ACT]VERB[GERUND]/NOUN`	'faxing'

6.2 Annotation of compounds

Compounding is encoded in the annotation by use of a + sign. A preverb followed by a verb is treated as a compound in this respect, as well as a *NOUN+NOUN* or an *ADJ+NOUN* compound. Compounding is similar to derivation in that only the last part of the word can be subjected to inflectional suffixing and that the output category of the compound is determined by the last component. E.g.:

rákkoktél 'shrimp coctail'
`rák/NOUN+koktél/NOUN`
keresztüllövi 'he shoots it through'
`keresztül/PREV+lő/VERB<DEF>`

7 Pronouns and postpositions

7.1 Pronouns

In Hungarian a pronoun can substitute for any noun, adjective or numeral, as well as for adverbs. The inflection of pronouns, where applicable, conforms to the restrictions imposed by the inflectional features and the tree-structure discussed above. This enables us to avoid the use of 'pronoun' as a POS category, and use instead the category which the pronouns stand for.

Personal pronouns are nouns, but they are subject to the following restrictions: their *POSS* feature must be negative and their *PERS* feature has to be specified. Otherwise, the *PERS* feature can combine with any other features (*PLUR*, *ANP*, *CAS*). E.g.:

ti 'you.PL'
`ti/NOUN<PERS<2>><PLUR>`
titeket 'you.PL.ACC'
`ti/NOUN<PERS<2>><PLUR><CAS<ACC>>`

Possessive pronouns are personal pronouns with a possessed feature, thus they carry the `ANP` feature as well. Examples include:

tiétek 'yours'
`ti/NOUN<PERS<2>><PLUR><ANP>`
tieteknek 'to/for yours'
`ti/NOUN<PERS<2>><PLUR><ANP><CAS<DAT>>`

The anaphoric possessive can be repeated as shown in the next example:

enyémé 'that of my something'
`én/NOUN<PERS<1>><ANP<ANP>>`

The above properties are shared by other pronouns including demonstrative, reflexive, relative, interrogative pronouns. The inflection of adjectival and numeral pronouns resemble to that of adjectives and numerals respectively, i.e. they are tagged as `ADJ` and `NUM` and take the usual inflections.

7.2 Postpositions

The function of postpositions is the same as that of case-suffixes, although some differences have to be noted. One major difference is that postpositions are separate words and, as such, have their own annotation. Furthermore, a number of postpositions can take the *PERS* feature and as their syntactic distribution (function) is the same as that of personal pronouns, these inflected postpositons will be annotated as nouns. In this case the *POSTP* feature of the tree also takes the positive value and the name of the relevant postposition has to be specified in the annotation as well[6]:

mellettetek 'next to you.PL'
`ti/NOUN<POSTP<MELLETT>><PERS<2>><PLUR>`

If the *POSTP* feature is positive, the *CAS*, *ANP* and *FAM* features have to be negative. Uninflected postpositions have the characteristics of a main POS category in that they can, for example, undergo derivation. Examples are:

mellett 'next to'
`mellett/POSTP`
mellettetek next to you.PL
`ti/NOUN<POSTP<MELLETT>><PERS<2>><PLUR>`
mellettiekben 'in those that are next to'
`mellett/POSTP[ATTRIB]/ADJ<PLUR><CAS<INE>>`

8 Derivational morphemes

The full list of derivational morphemes can be seen in Table 5. The output tag is followed by an (approximate) English name of the suffix and an allomorph. The input and output categories of the suffix are also indicated.

[6]The full list of tags that can be dominated by a *POSTP* tag can be seen in Table 7.2.

9 Comparison with other systems

The annotation system described in this document is independent of the implementation and the technical details of the morphological analysis. As such it is especially suitable to act as a common ground when comparing different formalisms.

While designng our system we examined the MSD coding system[1], which is positional, i.e. it has fixed positions for each morphosyntactic property and these positions can be either filled in or left empty. An MSD code is not suited to describe derivations, it deals only with inflectional suffixing. The mapping between the two systems is ambiguous, but we designed our annotation system in a way that it should contain at least as much information as the MSD system.

Bibliography

[1] T. Erjavec and M. Monachini. Specifications and notation for lexicon encoding. Technical report, Copernicus Project 106 MULTEXT-East, December 1997.

[2] András Kornai. A főnévi csoport egyeztetése. In Telegdi and Kiefer, editors, *Általános Nyelvészeti Tanulmányok*, XVII. Akadémiai Kiadó, Budapest, 1989.

[3] Viktor Trón. Attribútum-érték struktúrák. In László Kálmán, Viktor Trón, and Károly Varasdi, editors, *Lexikalista elméletek a nyelvészetben*. Tinta Könyvkiadó, Budapest, 2002.

number:	singular plural	(*sógor*) „simple" (*sógor-ok*) familiáris birtokos (*sógor-ék*)	<-PLUR> <+PLUR<-FAM>> <+PLUR<+FAM>>
possessor:	none overt possessor	 person: 1st (*sógor-om*) 2nd (*sógor-od*) 3rd (*sógor-a*) number: singular (*sógor-ai*) plural (*sógor-uk*)	<-POSS> <+POSS<+1><-2>> <+POSS<-1><+2>> <+POSS<-1><-2>> <+POSS<-PLUR>> <+POSS<+PLUR>>
possessed:	none overt possessed	 number singular (*sógor-é*) plural (*sógor-éi*)	<-ANP> <+ANP<-PLUR>> <+ANP<+PLUR>>
case:	„none" overt, one of 16 cases:	NOM (*sógor*) ACC (*sógort*) DAT (*sógor-nak*) INS (*sógor-ral*) CAU (*sógor-ért*) TRA (*sógor-rá*) SUE (*sógor-on*) SBL (*sógor-ra*) DEL (*sógor-ról*) INE (*sógor-ban*) ELA (*sógor-ból*) ILL (*sógor-ba*) ADE (*sógor-nál*) ALL (*sógor-hoz*) ABL (*sógor-tól*) TER (*sógor-ig*) FOR (*sógor-ként*)	<-CAS> <+CAS<+ACC>> <+CAS<+DAT>> <+CAS<+INS>> <+CAS<+CAU>> <+CAS<+TRA>> <+CAS<+SUE>> <+CAS<+SBL>> <+CAS<+DEL>> <+CAS<+INE>> <+CAS<+EAL>> <+CAS<+ILL>> <+CAS<+ADE>> <+CAS<+ALL>> <+CAS<+ABL>> <+CAS<+TER>> <+CAS<+FOR>>

Table 2: Inflectional features of nouns

modality:	none		< −MODAL>
		modal (*futhat*)	< +MODAL>
mood:	conjunctive		<−SUBJUNC><−COND>
		subjunctive/imperative	
		(no tense)	< +SUBJUNC>
		conditional	<+COND>
tense:	present		<−PAST><−FUT>
	past [5]		<+PAST>
	future		
	(only for the copula '*van*')		<+FUT>
number/person:	subject person		
		1st (*futok*)	<+PERS<+1><−2>>
		1st (*várlak*)	
		with 2nd person object	<+PERS<+1<+OBJ<+2><−2>>
		2nd (*futsz*)	<+PERS<−1><+2>>
		3rd (*fut*)	<+PERS<−1><−2>>
	subject number		
		singular (*fut*)	<−PLUR>
		plural (*futnak*)	<+PLUR>
definiteness	indefinite	(*lát*)	<−DEF>
	definite	(*látja*)	<+DEF>

Table 3: Inflectional features of verbs

ALÁ	(to) under X
ALATT	under X
ALÓL	from under X
ÁLTAL	by X, by way of X
ELÉ	before X, in front of X
ELÉB	before X, in front of X (archaic)
ELLEN	against X
ELLEN	contrary to X
ELŐL	from (in front of) X
ELŐTT	before X, in front of X
FELÉ	towards X
FELETT	above X, over X
FELŐL	from (the direction of) X, as for X
FELÜL	from (above/over) X
FÖLÉ	above X, over X
FÖLIBE	above X, over X (archaic)
FÖLÖTT	above X, over X
FÖLÜL	from (above/over) X
HELYETT	instead of X
IRÁNT	person marking with infixing
KÖRÉ	(to) around X
KÖRÖTT	around X
KÖRÜL	around X
KÖRÜLÖTT	around X
KÖZÉ	to (between many, among many) X
KÖZIBÉ	to (between many, among many) (archaic)
KÖZÖTT	between X, among X
KÖZT	between X, among X
KÖZÜL	out of X, from among X
LÉT	these can have inflected demostrative forms
MELLÉ	to somewhere near X
MELLETT	beside X, by X, (somewhere) near X
MELLŐL	from somewhere near X
MIATT	because of X
MÖGÉ	(to) behind X
MÖGÖTT	behind X
MÖGÜL	from (behind) X
NÉLKÜL	without X
RÉSZ	as concerns X
RÉSZ	for X
SZÁM	for X (recipient)
SZERINT	according to X
UTÁN	after X

Table 4: List of features that can combine with the feature *PERS*

Table 5: Derivational morphemes

Tag	explanation	example	POS
FREQ	frequentative	gat	VERB → VERB
MEDIAL	medial	ódik	VERB → VERB
CAUS	causative	tat	VERB → VERB
PART	adverbial participle	va	VERB → ADV
PERF_PART	perfect adverbial participle	ván	VERB → ADV
IMPERF_PART	imperfect adjectival participle	ó	VERB → ADJ
FUT_PART	future adjectival participle	andó	VERB → ADJ
PERF_PART	perfect adjectival participle	ott	VERB → ADJ
NEG_PERF_PART	negative perfect adjectival participle	atlan	VERB → ADJ
GERUND	gerund	ás	VERB → NOUN
NEG_MODAL_PART	negative modal adjectival participle	hatatlan	VERB → ADJ
MODAL_PART	modal adjectival participle	ható	VERB → ADJ
REG_ACT	regular activity	kodik	NOUN → VERB
ABSTRACT	abstract	ság	NOUN → NOUN
MRS	mrs	né	NOUN → NOUN
DIMIN	diminutive	ka	NOUN → NOUN
ATTRIB	attributive	s	NOUN → ADJ
MET_ATTRIB	metonymical attributive	i	NOUN → ADJ
INAL_ATTRIB	inalienable attributive	jú	NOUN → ADJ
NEG_ATTRIB	negative attributive	talan	NOUN → ADJ
TYPE1	type1	szeru	NOUN → ADJ
TYPE2	type2	féle	NOUN → ADJ
TYPE3	type3	nemu	NOUN → ADJ
TYPE_RANK	type rank	rangú	NOUN → ADJ
NEG_ATTRIB2	negative attributive2	mentes	NOUN → ADJ
TYPE4	type4	fajta	NOUN → ADJ
LOC_INE	locative inessive	beli	NOUN → ADJ
QUANTITY	quantity	nyi	NOUN → NUM
ESS_FOR	essivus formalis	képpen	NOUN → ADV
COM	comitative	stul	NOUN → ADV
PERIOD1	period1	anként	NOUN → ADV
PERIOD2	period2	onta	NOUN → ADV
ACT	activity	oz	NOUN → VERB
ACT2	activity2	ol	NOUN → VERB
COMPAR	comparative	bb	ADJ → ADJ
SUPERLAT	superlative	leg-bb	ADJ → ADJ
SUPERSUPERLAT	supersuperlative	legesleg-bb	ADJ → ADJ
COMPAR_DESIGN	comparative designative	bbik	ADJ → ADJ
SUPERLAT_DESIGN	superlative designative	leg-bbik	ADJ → ADJ
SUPERSUPERLAT_DESIGN	supersuperlative designative	legesleg-bbik	ADJ → ADJ
MANNER	manner	lag	ADJ → ADV
MANNER	manner	an	ADJ → ADV
INTRANS_RESULT	intransitive resultative	odik/ul	ADJ → VERB
TRANS_RESULT	transitive resultative	ít	ADJ → VERB
MULTIPL-ITER	multiplicative iterative	szor	NUM → ADV
MULTIPL-ITER	multiplicative iterative	szoroz	NUM → VERB
ITER_ATTRIB	iterative attributive	szori	NUM → ADJ
MULTIPL_ATTRIB	multiplicative attributive	szoros	NUM → ADJ
MULTIPL	multiplicative	szorta	NUM → ADV
AGGREG	aggregative	an	NUM → ADV
FRACT	fractional	ad	NUM → NUM
ORD	ordinal	odik	NUM → NUM
DATE	date	odika	NUM → NOUN
ATTRIB	attributive	i	POSTP → ADJ