Universal Morphology for Old Hungarian

Eszter Simon
Research Institute for Linguistics,
Hungarian Academy of Sciences
Benczúr u. 33.
H-1068 Budapest, Hungary
simon.eszter@nytud.mta.hu

Veronika Vincze
MTA-SZTE Research Group
for Artificial Intelligence
Tisza Lajos krt. 103.
H-6720 Szeged, Hungary
vinczev@inf.u-szeged.hu

Abstract

This paper provides a description of the automatic conversion of the morphologically annotated part of the Old Hungarian Corpus. These texts are in the format of the Humor analyzer, which does not follow any international standards. Since standardization always facilitates future research, even for researchers who do not know the Old Hungarian language, we opted for mapping the Humor formalism to a widely used universal tagset, namely the Universal Dependencies framework. The benefits of using a shared tagset across languages enable interlingual comparisons from a theoretical point of view and also multilingual NLP applications can profit from a unified annotation scheme. In this paper, we report the adaptation of the Universal Dependencies morphological annotation scheme to Old Hungarian, and we discuss the most important theoretical linguistic issues that had to be resolved during the process. We focus on the linguistic phenomena typical of Old Hungarian that required special treatment and we offer solutions to them.

1 Introduction

There is a growing interest not only in the natural language processing (NLP) community, but even among theoretical and historical linguists for building and using databases of historical texts. High quality historical corpora enriched with some kinds of linguistic information and metadata can provide a fertile ground for theoretical investigations. Several databases of historical texts have recently been created for various Indo-European languages, such as the Penn-Helsinki Parsed Corpus of Middle English (Kroch and Taylor, 2000), the Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Britto, 2002), or the Welsh Prose corpus (Thomas et al., 2007) and for non-Indo-European languages as well, such as the Old Hungarian Corpus (Simon, 2014).

Historical corpora represent a rich source of data, but only if the relevant information is specified in a computationally interpretable and retrievable way. Moreover, following the current standardisation efforts allows for cross-lingual comparative studies, as well as for longitudinal investigations on language change. With the recent increase in the number of annotated corpora, it seems advisable to move towards a harmonized common framework and methodology. Standardization always facilitates future research – in this case even for researchers who do not know the Old Hungarian language.

Natural language processing activities in Hungary were not synchronized in the past, hence similar resources were developed in parallel at different locations. As a consequence, there are two morphological analyzers for Hungarian: Hunmorph (Trón et al., 2005) and Humor (Novák, 2003). The former one has not been maintained recently, while the latter one is not freely available. Moreover, they use different formalisms, which share only one common property: they do not follow any international standards. For the morphological annotation of Old Hungarian texts, the Humor analyzer was used, thus all of the morphologically annotated texts are in a special format, which is hard to be interpreted for a non-Hungarian researcher. That is the reason behind the need of mapping the Humor formalism to a widely used universal tagset, for which we chose the Universal Dependencies (UD) framework.

The UD tagset and annotation scheme have just been adapted to Modern Hungarian (Vincze et al.,

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 118–127,
Berlin, Germany, August 11, 2016.

2016). In this paper, we report the adaptation of the morphological annotation scheme to Old Hungarian, and we discuss the most important theoretical linguistic issues that had to be resolved during the process. Section 2 briefly presents the international project Universal Dependencies and Morphology, then we summarize the part-of-speech (POS) tags and morphological features that are relevant for Old Hungarian. Section 3 gives a brief introduction of the Old Hungarian language and describes the morphologically annotated part of the Old Hungarian Corpus which has been converted into the UD tagset. Section 4 reports on our experiences in the conversion and discusses the specific linguistic issues concerning parts-of-speech and features. In Section 5, we contrast the annotation schemes developed for Old and Modern Hungarian. Conclusions and the planned future work end the paper in Section 6.

2 Universal Dependencies and Morphology

Universal Dependencies is an international project that aims at developing a unified annotation scheme for dependency syntax and morphology in a language-independent framework (Nivre, 2015). Currently (as of June 2016), there are annotated datasets available for 45 languages, including modern languages such as English, German, French, Hungarian and Irish, and old languages such as Ancient Greek, Coptic, Latin and Old Church Slavic, among others[1]. Datasets from all these languages apply the same tagsets at the morphological and syntactic levels and are annotated on the basis of the same linguistic principles, to the widest extent possible, however, in some cases, language-specific decisions had to be made. The benefits of using a shared tagset across languages enable interlingual comparisons from a theoretical point of view and also multilingual NLP applications can profit from a unified annotation scheme.

Standardized tagsets for both morphological and syntactic annotation have been constantly improved in the international NLP community. As for dependency syntax, Stanford dependencies is one of the most widely used tagsets (de Marneffe and Manning, 2008). For morphology, the MSD coding system was developed for a bunch of Eastern European languages including Hungarian (Erjavec, 2012). Interset functions as an interlingua for different morphological tagsets and it enables the conversion of different tagsets to the same morphological representation (Zeman, 2008). Rambow et al. (2006) defined a multilingual tagset for POS tagging and parsing, while McDonald and Nivre (2007) identified eight POS tags based on data from the CoNLL-2007 Shared Task (Nivre et al., 2007). Petrov et al. (2012) offered a tagset of 12 POS tags and applied this tagset to 22 languages.

POS	description
ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary
CONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	number
PART	particle
PRON	nominal pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
VERB	verb
X	other

Table 1: POS tags for Old Hungarian.

Now, Universal Dependencies is the latest standardized tagset that we are aware of. In its current form, morphological information is encoded in the form of POS tags and feature–value pairs. There is a fixed set of universal POS tags without the possibility of introducing new members, but features and values can have language-specific additions if needed. Features are divided into the categories lexical features and inflectional features. Lexical features are features that are characteristics of the lemmas rather than the word forms, whereas inflectional features are those that are characteristics of the word forms. Both lexical and inflectional features can have layered features: some features are marked more than once on the same word, e.g. a Hungarian noun may denote its possessor's number as well as its own number. In this case, the `Number` feature has an added layer, `Number[psor]`.

As mentioned above, Universal Morphology

[1] http://universaldependencies.org

annotates words with POS information and morphological features. Tables 1 and 2 summarize the POS tags and morphological features that are relevant for Old Hungarian, based on the annotation scheme created for Modern Hungarian, described at the UD website and in Vincze et al. (2016).

3 Old Hungarian

The Old Hungarian era lasted from 896 to 1526, the year of the occupation of the major part of the Hungarian Kingdom by the Ottoman Empire. The first part of this period (between 896–1350), documented by linguistic fragments and short coherent texts, is called the Early Old Hungarian period. The Late Old Hungarian period between 1350–1526 is the period of codices.

The Old Hungarian Corpus (Simon, 2014) contains all codices from the Late Old Hungarian period and several minor texts from the Early Old Hungarian period in their original orthographic form. Because of the heterogeneity of the Old Hungarian orthographic system, the original tokens had to be transcribed into their modernized form during a normalization step (for more details, see Oravecz et al. (2010)). Twelve of 47 codices have been normalized so far, and five of them have been morphologically analyzed and disambiguated.

The five codices are (in the order of the year of their writing/translation): Jókai Codex (after 1372/around 1448), Munich Codex (1466), Festetics Codex (before 1494), Guary Codex (before 1495) and Booklet on the Dignity of the Apostles (1521). These codices contain legends of saints, prayers, psalms, Bible translations and religious readings.

The Humor morphological analyzer was originally developed for Modern Hungarian and later it was extended to be capable of analyzing words containing morphological constructions, suffixes, paradigms and stems that were used in Old Hungarian but no longer exist in Modern Hungarian (Novák et al., 2013). Since the analyzer generates all potential morphological analyses for each token, a disambiguation step is required to select the most appropriate analysis. For this purpose, an HMM-based trigram tagger, PurePos (Orosz and Novák, 2012) was used, whose output was manually validated and corrected. This is the source data of the present conversion process, which contains 158,746 tokens altogether.

4 Language-specific extensions

Since the time interval of the Old Hungarian period is more than 600 years, several linguistic phenomena were in permanent change during this period. That is one of the reasons behind the heterogeneity of Old Hungarian texts. For instance, the progress in which postpositions became verbal particles or adverbs roots back to the Proto-Hungarian period and lasts even in the Modern Hungarian era, thus making a decision on their POS tag is far from trivial (discussed in more detail in Section 4.2). Such issues posed several problems during the conversion process, which are detailed in this section.

In examples, throughout the section, the relevant parts are emboldened. As a morphological description, we apply and follow the standard Leipzig Glossing Rules. The source of the example is provided in brackets after the translation. If the example is part of the Bible, the translation is copied from the King James Bible, and its biblical locus (book, chapter, verse) is also provided.

First, we discuss general issues of the conversion, then we illustrate specific cases that are relevant to only some or only one POS. Finally, challenges concerning morphological features are summed up.

4.1 General issues

Derivations changing part-of-speech

Hungarian has a great number of derivational suffixes, some of which change the POS of the word. These may derive – among others – verbs from nouns, e.g. *fül* (‘ear’) ∼ *fülel* (‘listen carefully’); nouns from adjectives, e.g. *vad* (‘wild’) ∼ *vadság* (‘wildness’); adjectives from nouns, e.g. *hold* (‘moon’) ∼ *holdbeli* (‘located on the moon’); or adverbs from adjectives, e.g. *víg* (‘merry’) ∼ *vígan* (‘merrily’) (for more details, see Törkenczy (2005)). They are formed either with a non-harmonic suffix or with harmonic two- or more-form suffixes, which are added to the stem. The choice of the appropriate harmonic variant is determined by vowel harmony (see below).

Hungarian derivational suffixes are denoted by the Humor morphological analyzer, but the UD formalism takes into account only the POS of the derived form and does not note the root and the derivational steps during which the final word form was created. During the conversion, POSs of words containing derivational suffixes

Feature	Description	POS
PronType	type of pronouns	ADV,DET,PRON
NumType	type of numerals	ADJ,ADV,DET,NUM
Reflex	reflexivity	PRON
Poss	possessive pronouns	PRON
Number	number	ADJ,ADV,AUX,NOUN,NUM,PRON,PROPN,VERB
Number[psor]	number of possessor	ADJ,NOUN,NUM,PRON,PROPN
Number[psed]	number of possessed	ADJ,NOUN,NUM,PRON,PROPN
Person	person	ADJ,ADV,AUX,PRON,VERB
Person[psor]	person of possessor	ADJ,NOUN,NUM,PRON,PROPN
Case	case	ADJ,NOUN,NUM,PRON,PROPN
Definite	definiteness	DET,VERB
Degree	degree	ADJ,ADV,NUM
VerbForm	form of the verb	ADJ,ADV,VERB
Mood	mood	AUX,VERB
Tense	tense	AUX,VERB
Aspect	aspect	ADJ,VERB
Voice	voice	ADJ,VERB

Table 2: Morphological features for Old Hungarian.

which do not change the lexical category were left unchanged, while POS-changing suffixes caused several difficulties. In addition to changing the POS, the lemma had also to be changed.

In the case of POSs which cannot be inflected, the full normalized word form can stand for the lemma as well. However, in those cases when the derived form may be inflected (verbs, nouns, adjectives), the lemma and the normalized form are not interchangeable. Thus the new lemma has to be generated from the old lemma and the harmonized form of the derivational suffix. Moreover, there are several irregular stems which may be changed before the derivational suffix, thus the converter must be capable to deal with them. The irregular stems occurring in the current version of the corpus are fully covered by the rules of the converter, but new stems may appear when expanding the corpus with new sources. Lemmas coming from the Humor morphological analyzer can be preserved in the 10th column of the CoNLL-U format, which is dedicated to any other annotation.

Allomorphs

In Hungarian, most suffixes harmonize with the stem they are attached to, which means that most suffixes exist in two or three alternative forms differing in the suffix vowel, and the selection of the suffix alternant is determined by the stem vowel(s). This phenomenon is known as vowel harmony, whose roots probably go back to the Proto-Uralic language, thus it exists in the Old Hungarian language as well.

There are several alternants in the Old Hungarian language which do not exist in Modern Hungarian and which therefore have specific markings in the formalism of Humor. An example of this phenomenon is the allomorph *-i*. In many cases, it is difficult or even impossible to decide whether it is the 3rd person singular form of the possessive suffix, or whether it marks the plurality of the possessed noun. For instance, the form *ÿgeretÿth* can be normalized either as *ígéret-é-t* ('promise-POSS.3SG-ACC'), or as *ígéret-e-i-t* ('promise-POSS.3SG-PL-ACC'). These forms get the morphological code N.PxS3=i.Acc or N.PxS3.Pl=i.Acc in the Humor formalism. However, these phenomena cannot be marked in the framework of UD, therefore they have been converted into the same feature–value pair as the corresponding Modern Hungarian suffix, without marking the surface form of the suffix. Since the CoNLL-U format of UD allows us to keep the original language-specific POS tags and morphological features, these kinds of information will not be lost.

4.2 Issues concerning parts-of-speech

Pronouns

In UD, only pronouns that substitute nouns are assigned the POS tag PRON, all the other pronouns are tagged according to the POS they stand for in the context. However, in the Old Hungarian Corpus, all pronouns – even those substituting other parts-of-speech – are tagged as pronouns. While converting the data, we could exploit the fact that pronouns inflected for case can only substitute nouns, compare the examples below:

	ilyetén	könyörgés-ek-et
(1)	such	prayer-PL-ACC
	'such prayers' (Kazinczy C. 26r)	

	soha	**ilyetén-t**	nem	ten-ni
(2)	never	such-ACC	not	do-INF
	'such thing never to do' (Jókai C. 107)			

Thus, inflected pronouns were automatically tagged as PRON. Words that were originally tagged as pronouns and occurred in the nominative case (i.e. they were not inflected) were assigned their UD POS tags with the help of lexical support: we defined lists for those pronouns and determined their UD POS tag manually. For instance, in Example 1, *ilyetén* was tagged as ADJ. These lists were then used in the automatic conversion process.

Postpositions

Some of the prepositional meanings found in other languages such as English are expressed in Hungarian by postpositions (Example 3) and case endings (Example 4). Hegedűs (2014) claims that there is historical evidence that the only difference between postpositions and case suffixes is that suffixes are monosyllabic and most of them show vowel harmony with the stem they are attached to. Syntactically, the two groups behave largely identically in Modern Hungarian.

	ház-a	**fölött**
(3)	house-POSS.3SG	above
	'above his house' (Festetics C. 57)	

	ház-á-**ba**
(4)	house-POSS.3SG-ILL
	'into his house' (Jókai C. 88)

Similarly to the forms of pronouns inflected for case (Example 5), some postpositions may form postpositional pronominal forms (Example 6). The former word forms can be regarded as a combination of a case marker and a marker for person and number, while the latter ones consist of a postposition plus the regular person/number endings.

	nek-em
(5)	DAT-1SG
	'to me' (Festetics C. 54)

	ellen-em
(6)	against-1SG
	'against me' (Jókai C. 103)

In the Old Hungarian Corpus, however, these suffixes are analyzed as possessive endings, which is also a valid approach. Some of the Old Hungarian postpositions can appear in a structure that is analogous to the possessive construction (for more details on possessive constructions, see Section 4.3). Similarly to how the possessor can appear in dative case, the complement of some postpositions can also be in dative case, while a possessedness marker may appear on the postposition (Hegedűs, 2014), compare the examples below:

	halál-a	**után**
(7)	death-POSS.3SG	after
	'after his death' (Vienna C. 4)	

	halál-od-nak	**után-a**
(8)	death-POSS.2SG-DAT	after-POSS
	'after your death' (Bod C. 14r)	

Since inflected pronouns and inflected postpositions behave in a similar way, it can be argued that these endings are only markers of person and number, without referring to possession. In the UD morphology, we analyze both of them as personal pronouns as they can substitute inflected nouns, and assign them the features Person and Number, without any reference to possession.

Complex verb forms

According to the description on the UD website, auxiliaries express grammatical distinctions not carried by the lexical verb, thus the lexical verb and the auxiliary together bear all suffixes. In this sense, there are four auxiliaries in Old Hungarian (*vala, volt, volna, legyen*), which are parts of the Old Hungarian complex verb forms. In Hungarian, a conjugated verb form consists of the stem

plus two inflectional slots, i.e. positions where inflectional suffixes can occur. The first of these suffix positions is that of tense/mood and the second one is that of person/number. This is the reason behind the need for complex verb forms, thus there is insufficient place in one inflected word form for expressing tense and mood at the same time. Therefore, one of the tense and mood markers has to be 'out-sourced' to an auxiliary, while agreement and definiteness markers stay on the lexical verb.

There are four complex verb forms in Old Hungarian: past continuous, past perfect, past conditional, and past subjunctive. With the only exception of past conditional, all of them are extinct from the Modern Hungarian language.

The past continuous and the past conditional constructions have a version in which the auxiliary also bears an agreement marker, as in Examples 9 and 10:

(9)
tart-om **val-ék**
keep-1SG.DEF be-IPFV.1SG
'I was keeping (them)'
(Munich C. 103vb)

(10)
ír-t-am **vol-nék**
write-PST-1SG be-COND.1SG
'I would have written' (Bod C. 15r)

In these cases, `Person` and `Number` features of both the lexical verb and the auxiliary have the same value. In the cases where the auxiliary does not carry any grammatical distinctions, but the tense or mood suffixes, `Person`, `Number`, `Voice` and `Definite` features remain underspecified.

Verbal particles

Hungarian verbs often have particles, which appear pre-verbally in neutral Hungarian sentences. In these cases, they are attached to the beginning of the verb, thus they constitute one token with the verb (Example 11). However, there are several cases when particles become separated from the verb and actually appear after the verb. For example, if another word or group of words is the focus in the sentence, the particle obligatorily follows the verb (Example 12).

(11)
ki-tisztul-ok nagy vétés-ből
out-purge-1SG big sin-ELA
'I am purged from big sin'
(Festetics C. 11)

(12)
sok-ak-at **hagy-t-am** **el**
many-PL-ACC leave-PST-1SG away
'I left many' (Könyvecse 18v)

If the verbal particle immediately precedes the verb, its code is attached to that of the verb in the Humor formalism. Since the verbal particle + verb construction is treated as one unit, only one POS tag can be assigned to it, which is `VERB`.

In cases when the particle is separated from the verb, the particle itself must have its own POS tag. According to the UD description, however, not all function words that are traditionally called particles automatically qualify for the `PART` tag, but they may be adpositions or adverbs by origin, therefore should be tagged as `ADP` or `ADV`, respectively.

The state and origin of verbal particles are constantly disputed even in Modern Hungarian. For example, D. Mátai (1992) claims that they developed from spatial adverbs, while Hegedűs (2014) proposes that they all go back to spatial postpositions with a lative (mostly goal) meaning.

The oldest particles are *meg* 'back', *ki* 'out', *le* 'down', *el* 'away', *be* 'into', *fel* 'up'. They are telicizing elements with often little spatial meaning left due to semantic bleaching. However, since they have not been fully grammaticalized, they have preserved some spatial meaning, and as a result we cannot treat them as regular particles.

In addition to the oldest particles, several new ones were born during the Old Hungarian period. According to the theory of Hegedűs (2014), all of them go back to, and are grammaticalized from postpositions, therefore we tagged them as `ADP`.

Adverbial participles

Old Hungarian has three types of adverbial participles, which are formed with one of the harmonising two-form suffixes: *-ván/-vén*, *-va/-ve*, and *-atta/-ette*. In the UD formalism, they all have the `VerbForm=Trans` feature–value pair, since they are transgressives, i.e. non-finite verb forms that share properties of verbs and adverbs.

While *-ván/-vén* adverbial participles do not agree, participles with *-va/-ve* can optionally agree with their subject (Examples 13 and 14), and participles with *-atta/-ette* ending obligatorily agree with their subject, see Example 15.

(13)
hal-va lel-ik val-a
dead-PART find-3PL.DEF be-PST
'they found him dead' (Guary C. 103)

(14)
mi **alu-vánk**
we sleep-PART.1PL
'while we slept'
(Munich C. 35vb; Matthew 28,13)

(15)
míg ő **beszéll-ette**
while he speak-PART.3SG
'while he yet spake'
(Munich C. 81vb; Luke 22,47)

While some of the Old Hungarian non-finites do agree with their subject, none of them distinguish the definite and indefinite conjugation like finite clauses do. Moreover, they do not bear temporal, mood, and aspect suffixes, thus in this sense their agreement paradigm can be said to be defective. Therefore, they can optionally get the `Person` and `Number` features in UD besides the `VerbForm=Trans` feature–value pair.

4.3 Issues concerning features

Definiteness of the verb

As a special type of agreement, Hungarian verbs also mark the definiteness of their objects. In other words, the form of the verb changes when the definiteness of the object also changes (Törkenczy, 2005). Proper nouns and noun phrases with a definite article are prototypical examples of definite objects while bare nouns and noun phrases with an indefinite article are indefinite objects. Compare:

(16)
lát-á az ház-at
see-IPFV.3SG.DEF the house-ACC
'he saw the house' (Kazinczy C. 13r)

(17)
lát-a álm-ot
see-IPFV.3SG.INDEF dream-ACC
'he had a dream' (Vienna C. 73)

As can be seen in Examples 16 and 17, the two verb forms differ only in one accent, more precisely, in the definite form there is an accented *a*, but in the indefinite form, there is no accent on the last vowel. However, due to the lack of standardized orthography and spelling conventions in the Old Hungarian period, the very same words can be spelled completely differently on the one hand, and different words can be spelled in the same way on the other hand, especially when no diacritics are used. Thus, we could encounter cases when it was impossible to decide whether the definite or the indefinite form of the verb was meant to be used, e.g. *lata* could be *láta* (the indefinite form) as well as *látá* (the definite form). For these cases, it seemed necessary to add another possible value of the `Definite` feature: the value `Underspecified` denotes that the definiteness of the verb cannot be figured out and it leaves this feature under-specified.

Possessive constructions

The possessor in Hungarian possessive constructions can have two different surface forms both in Old and Modern Hungarian, without any difference in meaning (similar to the English constructions *the boy's dog* and *the dog of the boy*). That is, both of the following examples are widely used:

(18)
Jézus tanítvány-a
Jesus disciple-POSS.3SG
'Jesus's disciple'
(Munich C. 35rb; Matthew 27,57)

(19)
Jézus-nak nev-é-be
Jesus-DAT name-POSS.3SG-ILL
'in the name of Jesus' (Booklet 16r)

The first (unmarked) form coincides with the nominative case whereas the second (marked) form coincides with the dative form of the noun, cf.:

(20)
mond-á **Jézus-nak**
say-IPFV.3SG.DEF Jesus-DAT
'said unto Jesus'
(Munich C. 23rb; Matthew 17,4)

According to the UD guidelines for Modern Hungarian, the case of the unmarked possessor is nominative, that is, a nominative possessor is not distinguished from the subject. However, the marked possessor is labeled differently from the dative argument, bearing a genitive label. In the original version of the Old Hungarian Corpus, a distinction was made in all of the cases, and the labels Nom, Dat, Nom_Gen and Dat_Gen are used for the subject, indirect object, nominative possessor and dative possessor, respectively.

Here, we voted for not making a distinction of the surface cases at the level of morphology. Hence, we annotated the unmarked possessor with the nominative case and the marked possessor with the dative case. On the other hand, the syntactic annotations of these should differ from each other, that is, the distinction will be made at the level of syntax. Table 4.3 summarizes these distinctions.

Example	Translation	UD for MH	OH original	UD for OH
a fiú kutyája	the boy's dog	Nom	Nom_Gen	Nom
a fiú játszott	the boy was playing	Nom	Nom	Nom
a fiúnak a kutyája	the dog of the boy	Gen	Dat_Gen	Dat
a fiúnak adta a könyvet	he gave the book to the boy	Dat	Dat	Dat

Table 3: Morphological features for possessors (MH: Modern Hungarian, OH: Old Hungarian).

5 Differences between Old and Modern Hungarian

In this section, we briefly contrast the annotation schemes for Old and Modern Hungarian, and we highlight the most important differences.

In Old Hungarian, there were more tenses and verb forms in use than in Modern Hungarian (see Section 4.2). Hence, more feature combinations are possible in Old Hungarian. Certain forms of adverbial participles agreed with the subject in Old Hungarian, however, this phenomenon is extinct now (cf. Section 4.2). For this reason, adverbial participles can have the features `Number` and `Person` in Old Hungarian but not in Modern Hungarian.

The verbal particle *meg* originates from a postposition meaning 'behind'. However, in Modern Hungarian, *meg* totally lost this shade of meaning and now is only used as a particle that perfectivizes the meaning of the verb it is attached to. Due to this historical change, *meg* is tagged as `PART` in Modern Hungarian but as `ADP` in Old Hungarian.

In Old Hungarian, ordinal and fractal numbers are not distinguished from each other, that is, the word form *harm-ad* ('three-DERIV.SFX') can mean 'a third part of something' and 'the third one' as well. However, in Modern Hungarian, it can only have the first meaning, the latter one is expressed by the word form *harm-ad-ik* ('three-DERIV.SFX-DES'). As a consequence, fractal numbers occur only in Modern Hungarian but not in Old Hungarian.

There are also differences concerning the marking of possessors. As discussed above in Section 4.3, the Old Hungarian UD annotation scheme makes use of only the labels `Nom` and `Dat`, regardless of whether the noun is used as a possessor or not. However, the morphological annotation of the UD treebank for Modern Hungarian was converted from the Szeged Treebank (Csendes et al., 2005), which makes a distinction between dative possessors and indirect objects (both ending in a dative suffix), thus the distinction was kept in the UD treebank as well. It should be noted, however, that it is not historical changes that led to this distinction: the annotation principles of the two treebanks are responsible for this divergence.

Due to the orthographic features of codices, the value `Underspecified` had to be added to the `Definite` feature for verbs, which is not present in Modern Hungarian (cf. Section 4.3). Nevertheless, this feature value might be of use in Modern Hungarian too: for instance, social media users tend to write their posts without accents, which might also yield ambiguous word forms. Thus, should social media texts be included in the Modern Hungarian UD treebank in the future, this feature value might be exploited there as well.

As can be seen, in some cases, Old Hungarian had a richer set of morphological processes (for instance, verbal conjugation), but in other cases, Modern Hungarian has developed some more morphological distinctions (like that of ordinal and fractal numbers). Thus, both additions and losses occurred in Hungarian morphology from a historical perspective. Later on, we intend to investigate whether this is true for syntax as well: we would like to adapt the UD annotation guidelines to Old Hungarian and see the syntactic differences between Old and Modern Hungarian.

6 Conclusions and future work

In this paper, we reported the automatic conversion of the morphological annotation of the Old Hungarian Corpus to the international standard framework of Universal Dependencies and Morphology. We presented the linguistic phenomena typical of Old Hungarian that required special treatment and we offered solutions to them. The detailed description of the Old Hungarian morphology has been made publicly available, together with the converted corpus[2]. Later on, we intend to adapt the Modern Hungarian UD depen-

[2]http://oldhungariancorpus.nytud.hu/

dency tagset and annotation principles to Old Hungarian as well. After that, we are planning to add syntactic annotation to the corpus and publish it at the UD website[3], together with the adapted dependency labels and their detailed description.

Currently, additional texts from the Old Hungarian period are being digitized and normalized, also, morphological annotation is being added to them. These texts will then be standardized according to the UD morphology on the basis of the conversion rules developed in this paper and thus, the dataset of Old Hungarian texts with UD morphology will be expanded too.

Finally, it should be noted that the Hungarian NLP community is currently implementing a new morphological analyzer, which is planned to provide output in different formalisms, one of which will be the UD morphology. We are confident that our corpus and the above-mentioned morphological analyzer can contribute to the more effective and faster processing of Old Hungarian texts.

Acknowledgments

The research reported in the paper was conducted with the support of the Hungarian Scientific Research Fund (OTKA) grant #112057. We thank the anonymous reviewers for their comments.

References

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged TreeBank. In Václav Matousek, Pavel Mautner, and Tomás Pavelka, editors, *Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005*, Lecture Notes in Computer Science, pages 123–132, Berlin / Heidelberg, September. Springer.

Mária D. Mátai. 1992. Az igekötők [Particles]. In Loránd Benkő, editor, *A magyar nyelv történeti nyelvtana II/1. A kései ómagyar kor. Morfematika [Historical grammar of the Hungarian language. The Late Old Hungarian period. Morphology]*, pages 662–695. Akadémiai Kiadó, Budapest.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford dependencies manual. Technical report, Stanford University.

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

[3] As currently there is no dependency annotation available for the Old Hungarian Corpus, it is not officially listed among the UD treebanks on the UD website.

Charlotte Galves and Helena Britto. 2002. The Tycho Brahe Corpus of Historical Portuguese. Online publication.

Veronika Hegedűs. 2014. The cyclical development of Ps in Hungarian. In É. Kiss, Katalin, editor, *The Evolution of Functional Left Peripheries in Hungarian Syntax*, pages 122–147. Oxford University Press.

Anthony Kroch and Ann Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). CD-ROM.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 122–131.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Attila Novák, György Orosz, and Nóra Wenszky. 2013. Morphological annotation of Old and Middle Hungarian corpora. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 43–48, Sofia, Bulgaria, August. Association for Computational Linguistics.

Attila Novák. 2003. Milyen a jó Humor? [What is good Humor like?]. In *Proceedings of the 1st Hungarian Computational Linguistics Conference*, pages 138–144, Szeged. SZTE.

Csaba Oravecz, Bálint Sass, and Eszter Simon. 2010. Semi-automatic Normalization of Old Hungarian Codices. In *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, pages 55–60, Lisbon, Portugal. Faculty of Science, University of Lisbon.

György Orosz and Attila Novák. 2012. PurePos: An Open Source Morphological Disambiguator. In *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*, pages 53–63.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*, May.

Owen Rambow, Bonnie Dorr, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith J. Miller, Teruko Mitamura,

Reeder, Florence, and Advaith Siddharthan. 2006. Parallel syntactic annotation of multiple languages. In *Proceedings of LREC*, May.

Eszter Simon. 2014. Corpus building from Old Hungarian codices. In É. Kiss, Katalin, editor, *The Evolution of Functional Left Peripheries in Hungarian Syntax*, pages 224–236. Oxford University Press.

Peter Wynn Thomas, D. Mark Smith, and Diana Luft. 2007. Rhyddiaith Gymraeg 1350-1425.

Miklós Törkenczy. 2005. *Practical Hungarian Grammar*. Corvina, Budapest.

Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. 2005. Hunmorph: Open source word analysis. In *Proceedings of the ACL Workshop on Software*, pages 77–85, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Veronika Vincze, Richárd Farkas, Katalin Ilona Simkó, Zsolt Szántó, and Viktor Varga. 2016. Univerzális morfológia és dependencia magyar nyelvre [Universal Morphology and Dependencies for Hungarian]. In *XII. Magyar Számítógépes Nyelvészeti Konferencia*.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.