Universal Dependencies and Morphology for Hungarian – and on the Price of Universality

Veronika Vincze[1,2], **Katalin Ilona Simkó**[1], **Zsolt Szántó**[1], **Richárd Farkas**[1]
[1]University of Szeged
Institute of Informatics
[2]MTA-SZTE Research Group on Artificial Intelligence
kata.simko@gmail.com
{vinczev,szantozs,rfarkas}@inf.u-szeged.hu

Abstract

In this paper, we present how the principles of universal dependencies and morphology have been adapted to Hungarian. We report the most challenging grammatical phenomena and our solutions to those. On the basis of the adapted guidelines, we have converted and manually corrected 1,800 sentences from the Szeged Treebank to universal dependency format. We also introduce experiments on this manually annotated corpus for evaluating automatic conversion and the added value of language-specific, i.e. non-universal, annotations. Our results reveal that converting to universal dependencies is not necessarily trivial, moreover, using language-specific morphological features may have an impact on overall performance.

1 Introduction

Morphological tagging and syntactic parsing are key components in most natural language processing (NLP) applications. Linguistic resources and parsers for morphological and syntactic analysis have been developed for several languages, see e.g. the shared tasks on morphologically rich languages (Seddah et al., 2013; Seddah et al., 2014). However, the comparison of results achieved for different languages is not straightforward as most languages and databases apply a unique tagset, moreover, they were annotated following different guidelines. In order to overcome these issues, the project Universal Dependencies and Morphology (UD) has recently been initiated within the NLP community (Nivre, 2015). The main goal of the UD project is to develop a "universal", i.e. a language-independent morphological and syntactic representation which can contribute to the implementation of multilingual morphological and syntactic parsers from a computational linguistic point of view. Furthermore, it can enhance studies on linguistic typology and contrastive linguistics.

From the viewpoint of syntactic parsing, the languages of the world are usually categorized according to their level of morphological richness (which is negatively correlated with configurationality). At one end, there is English, a strongly configurational language while there is Hungarian at the other end of the spectrum with rich morphology and free word order (Fraser et al., 2013). In this paper, we present how UD principles were adapted to Hungarian, with special emphasis on Hungarian-specific phenomena.

Hungarian is one of the prototypical morphologically rich languages thus our UD principles can provide important best practices for the universalization of other morphologically rich languages. The UD guidelines for Hungarian were motivated by both linguistic considerations and data-driven observations. We developed a converter from the existing Szeged Dependency Treebank (Vincze et al., 2010) to UD and manually corrected 1,800 sentences from the newspaper domain. The experiences gained during the converter development and during the manual correction could reinforce the linguistic guidelines. Moreover, the manually corrected gold standard corpus provides the opportunity for empirical evaluations like assessing the converter and comparing dependency parsers employing the original and the universal morphological representations. Thus, we evaluated the quality of the automatic conversion, which reveals that converting to universal dependencies is not necessarily trivial, at least for Hungarian. We also show that using different morphological tagsets may have an impact on overall parsing performance and utilizing language-specific, i.e. non-universal, information has a considerable

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 356–365,
Valencia, Spain, April 3-7, 2017.

added value at both the morphological and syntactic layers.

The chief contributions of the paper are the introduction of

- the universal morphology and dependency principles for Hungarian, leading to insights for other morphologically rich languages,
- empirical experiments on the upper bound of the accuracy of automatic conversion and pre-parsing,
- comparative evaluations for assessing the added value of language-specific information at the morphological and syntactic layers along with the interaction of these two.

2 Related Work

Standardized tagsets for both morphological and syntactic annotations have been constantly developed in the international NLP community. For instance, the MSD morphological coding system was developed for a set of Eastern European languages (Erjavec, 2012), within the MULTEXT-EAST project. Interset functions as an interlingua for several morphological coding systems, which can convert different tagsets to the same morphological representation (Zeman, 2008). There have also been some attempts to define a common set of parts-of-speech: Rambow et al. (2006) defined a multilingual tagset for part-of-speech (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.

Now, Universal Dependencies (UD) is the latest standardized tagset that we are aware of. UD is an international project that aims at developing a unified annotation scheme for dependency syntax and morphology in a language-independent framework (Nivre, 2015). Hungarian was among the first 10 languages of the project, participating also in the first official release in January 2015. In the latest release (Version 1.3, May 2016), there are annotated datasets available for 40 languages, including English, German, French, Hungarian and Irish, among others[1]. In these datasets, the very same tagsets are applied at the morphological and syntactic levels and texts are annotated on the basis of the same linguistic principles, to the widest extent possible.

The UD tagset encodes morphological information in the form of POS tags and feature–value pairs. As for syntactic information, each word is assigned to its parent word in the dependency tree and the grammatical function of the specific word is encoded in dependency labels. Dependency labels, POS tags and features are universal (i.e. there is a fixed set of them without the possibility of introducing new members), but values and dependency labels 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, Num[psor].

Up to now, several papers have been published on the general principles behind UD (Nivre, 2015; Nivre et al., 2016) or on specific treebanks. For instance, there are UD treebanks available for agglutinative languages such as Finnish (Haverinen et al., 2014; Pyysalo et al., 2015), Estonian (Muischnek et al., 2016) and Japanese (Tanaka et al., 2016), for Slavic languages (Zeman, 2015) and spoken Slovenian (Dobrovoljc and Nivre, 2016) and for Nordic languages such as Norwegian (Øvrelid and Hohle, 2016), Danish (Johannsen et al., 2014) and Swedish (Nivre, 2014), together with several other languages (Persian (Seraji et al., 2016) and Basque (Aranzabe et al., 2014), just to name a few). Recently, a further extension on the UD relations has been proposed: enhanced English dependencies are described in Schuster and Manning (2016).

Our UD principles introduced in this paper follow the central UD guidelines (Nivre, 2015) and we did our best to align with the existing guidelines for other morphologically rich languages as well. On the other hand, there are several Hungarian-specific phenomena that required changes and extensions of the original UD principles.

The only available manually annotated tree-

[1] http://universaldependencies.org/

bank for Hungarian is the Szeged Corpus (Csendes et al., 2004) and Szeged Dependency Treebank (Vincze et al., 2010). It contains approximately 82,000 sentences and 1.5 million tokens, all manually annotated for POS-tagging and constituency and dependency syntax. We developed an automatic tool that converts the morphological descriptions of the Szeged Corpus to universal morphology tags and the dependency trees of the Szeged Treebank to universal dependencies.

3 Universal Morphology for Hungarian

In this section, we present the morphological tagset applied to Hungarian.

When adapting the principles of Universal Morphology to Hungarian, we were able to automatically convert most of the morphological features used in the Szeged Treebank 2.5 (Vincze et al., 2014), which was based on MSD principles (Erjavec, 2012). However, we faced some problematic issues, which we will discuss in detail in this section. The details of universal morphological codeset of Hungarian are available on our website[2].

3.1 Possessive constructions

The possessor in Hungarian possessive constructions can have two different surface forms, without any difference in meaning: the possessor can be morphologically marked or not, just like the English constructions *the girl's doll* and *the doll of the girl*. Thus, both of the following possessive constructions are widely used:

(1) a szomszéd kertje
the neighbor garden-3SGPOSS
the neighbor's garden

(2) a szomszédnak a kertje
the neighbor-DAT the garden-3SGPOSS
the neighbor's garden

In Example 1, the possessor is not marked, i.e. it shares its form with the nominative form of the noun, however, in Example 2, the possessor is morphologically marked, sharing its from with the dative form of the noun. Nevertheless, the possessed is morphologically marked in both cases, which was a novelty in the UD project as the languages already included in the data do not mark the possessor on the possessed noun but use determiners for this purpose (cf. *my car* but *az autóm* (the car-1SGPOSS)). Moreover, the number of the possessed can be marked on the noun in elliptical constructions such as:

(3) Láttam az
see-PAST-2SGPOSS-ACC the
autódat , de a
car-2SG-POSS , but the
szomszédét nem .
neighbor-POSSD.SG-ACC not
I could see your car but not that of the neighbor.

(4) Láttam a
see-PAST-2SGPOSS-ACC the
gyerekeidet , de a
child-2SG-PL-POSS , but the
szomszédéit nem .
neighbor-POSSD.PL-ACC not
I could see your children but not those of the neighbor.

Hence, we had to introduce novel morphological features to mark the person and number features of the possessor on Hungarian nouns. Number denotes the number of the noun, Number[psor] and Person[psor] denote the number and person of the possessor, and Number[psed] denotes the number of the possessed. Below, there is a sample word annotated according to the Universal Morphology principles.

(5) házaimének
house-1SGPOSS-PL-POSSD.SG-DAT
to that of my houses
NOUN
Case=Dat|Number=Plur|Number[psed]=Sing
|Number[psor]=Sing|Person[psor]=1

3.2 Object-verb agreement

Another Hungarian-specific feature was the definiteness of the object. As a special type of agreement, the definiteness of their objects determines which paradigm of the verb is to be chosen. In other words, the form of the verb changes when the definiteness of the object also changes (Törkenczy, 2005). For instance, proper nouns and NPs with a definite article are typical examples of definite objects and trigger the objective form of the verb (see Example 6) while bare nouns and NPs with an indefinite article are indefinite objects

[2] http://rgai.inf.u-szeged.hu/project/nlp/research/msdkr/univmorph.html

(see Example 7) and trigger the subjective form of the verb. Second person objects also trigger a special form of the verb as listed in Example 8:

(6) Látom Pistit .
see-1SGOBJ Steve-ACC .
I can see Steve.

(7) Látok egy gyereket az udvaron .
see-1SGSUBJ a kid-ACC the yard-SUP .
I can see a kid in the yard.

(8) Látlak .
see-1SGOBJ2 .
I can see you.

In this way, the feature Definiteness needs to be applied to verbs in Hungarian, moreover, it has a language-specific feature due to the special form triggered by the second person objects. Thus, Definiteness has three possible values in Hungarian: `Definite`, `Indefinite`, `2`.

3.3 Determiners and pronouns

Determiners, pronouns and ordinal numbers also constituted a peculiarity. According to Hungarian grammatical traditions, ordinal numbers have been treated as numerals but in the universal morphology, they have to be annotated as adjectives. Thus, their POS tags were automatically converted to adjectives.

Demonstrative pronouns were also treated differently in the original annotation used in the Szeged Treebank and in universal morphology. While demonstrative pronouns *ez* and *az* are tagged as pronouns independently of their positions, in universal morphology such words occurring before an article should be tagged as a determiner (see Example 9) but when they are used as an NP, they should be tagged as a pronoun (see Example 10).

(9) Olvastam azt a könyvet .
read-PAST-1SGOBJ that-ACC the book-ACC .
I have read that book.

(10) Olvastam azt .
read-PAST-1SGOBJ that-ACC .
I have read that.

These cases were also automatically converted, following the universal morphology guidelines.

3.4 Verbal prefixes

In our original treebank, verbal particles that were spelt as a separate token had their own part-of-speech, i.e. verbal particle. According to the UD description however, not all function words that are traditionally called particles automatically qualify for the PART tag. They may be adpositions or adverbs by origin, therefore should be tagged ADP or ADV, respectively. Thus, we manually compiled a list that contained the original part-of-speech of words that were tagged as verbal prefixes, for instance, *el* "away" was treated as an adverb and *agyon* brain-SUP as a noun – the latter is usually used in phrases like *agyonüt* "kill someone by hitting on his head". Based on this list, we were able to automatically assign UD POS tags to verbal prefixes.

4 Universal Dependency in Hungarian

When adapting the universal dependency labels to Hungarian, we could find a one-to-one correspondence between the original labels of the Szeged Treebank and the UD labels only in most of the cases, and these labels could be automatically converted to the UD format, making use of the dependency and morphological annotations found in the original treebank. However, we encountered some problematic cases during conversion, which we will discuss below in detail. The details of universal dependency rules of Hungarian are available on our website[3].

4.1 Non-overt copulas

Traditionally, it is the verb that functions as the head of the clause in dependency grammars but in certain languages, there are verbless clauses where the predicate consists of a single nominal element (typically a noun or an adjective) at the surface level. The dependency analysis of such sentences may be problematic due to the lack of an overt verb. Some studies such as Polguère and Mel'čuk (2009) argue for a zero copula in such cases, especially when the copula is empty only in certain slots of the verbal paradigm. For instance, in Hungarian, the copula has its zero form only in the present tense, indicative mood, third person forms as shown in Examples 11-14:

(11) Present tense, indicative mood, Sg1:

[3] `http://rgai.inf.u-szeged.hu/dependency`

Én tanár vagyok .
I teacher be-1SG .

I am a teacher.

(12) Present tense, indicative mood, Sg3:

Ő tanár .
he teacher .

He is a teacher.

(13) Past tense, indicative mood, Sg3:

Ő tanár volt .
he teacher be-PAST-3SG.

He was a teacher.

(14) Present tense, imperative mood, Sg3:

Ő legyen tanár !
he be-IMP-3SG teacher !

He should be a teacher.

The original dependency analysis in the Szeged Treebank inserts a zero copula (VAN), i.e. a virtual node in the dependency tree, which functions as the head of the clause and the nominal predicate is attached to it. Figure 1 shows such an analysis of the sentence *E gondolat sem új* (this thought not new) "This thought is not novel at all".

Beside the function head analysis (i.e. where function words, e.g. the copula is the head), there is another approach to dependencies, namely, the content head analysis, where the head is a content word instead of a function word. In the latter case, the main grammatical relations can be found among content words and all the other function words are attached to the main structure. UD applies the content head analysis, which means that in copular constructions, the nominal element is the head and the copula (if present) is attached to it with a `cop` relation. In a similar way, the head of adpositional constructions is the noun and the adposition is attached to it.

Sentences with nominal predicates were automatically converted from the original treebank into the UD format: Figure 2 shows the UD analysis of the sentence found in Figure 1. Likewise, postpositional constructions were converted: the noun was treated as the head and the postposition was attached to it with a `case` label.

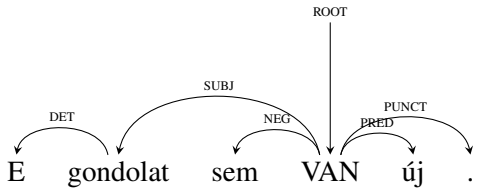

Figure 1: A function head analysis in the Szeged Dependency Treebank (*E gondolat sem VAN új* (this thought IS not new) "This thought is not novel at all").

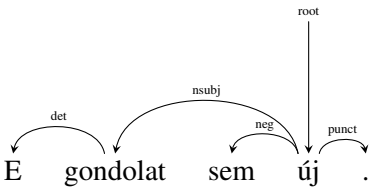

Figure 2: A content head analysis in the Hungarian UD treebank (*E gondolat sem új* (this thought not new) "This thought is not novel at all").

4.2 Subordinate clauses

Subordinate clauses proved also to be a problematic issue as UD principles make a sharp distinction among several types of subordinate clauses – e.g. clausal subject, clausal object, adverbial clause – in contract with the Szeged Dependency Treebank, which applies one single label for all types of subordinate clauses. Some types of subordinate clauses had a special label in the constituency version of the treebank hence their conversion was straightforward. In other cases, we could rely on manually constructed conversion rules but the resulting trees had to be corrected manually.

4.3 Multiword named entities

The UD treatment of multiword named entities required a Hungarian-specific solution. According to the UD principles, the first token of the multiword expressions should be marked as the head. However, in Hungarian, it is always the last element of the multiword expression that is inflected. Examples 15-16 demonstrate that the first element cannot be inflected, only the last one:

(15) Találkoztam Kovács Jánossal .
meet-PAST-1SG Kovács János-INSTR .

I met János Kovács[4].

[4]The standard order of person names is surname + first

(16) *Találkoztam Kováccsal János .
meet-PAST-1SG Kovács-INSTR János .
I met János Kovács.

Due to the above morphosyntactic facts, we marked the last token of multiword named entities as the head in the Hungarian UD treebank while all the other UD treebanks mark the first token as the head.

4.4 Dative forms

In Hungarian, nouns that bear the suffix *-nAk* can fulfill several grammatical roles in the sentence such as:

(17) indirect object:

Laci adott a
Leslie give-PAST-3SG the
barátjának egy almát .
friend-3SGPOSS-DAT an apple-ACC .

Leslie gave an apple to his friend.

(18) possessor:

Laci elvette a
Leslie take-PAST-3SGOBJ the
barátjának a
friend-3SGPOSS-DAT the
könyvét .
book-3SGPOSS-ACC .

Leslie took his friend's book.

(19) dativus ethicus:

Nekem nehogy eladd
I-DAT so.as.not.to sell-IMP-2SGOBJ
az autódat !
the car-2SGPOSS-ACC !

As for me, you should not sell you car.

(20) experiencer:

Nekem nagyon tetszett az
I-DAT very like-PAST-3SG the
előadás .
performance .

I really liked the performance.

(21) semantic subject:

Lacinak bocsánatot
Leslie-DAT apology-ACC
kellett kérnie a
must-PAST-3SG ask-INF-3SG the
barátjától .
friend-3SGPOSS-ABL .

Leslie had to apologize to his friend.

While these forms do not show any difference at the morphological level, they have very different roles at the syntactic and semantic levels. Thus we decided not to make any distinction in the morphological annotation but they should have different syntactic labels. Indirect objects are marked with the label *iobj*, possessors with the label *nmod:poss* and other occurrences with *nmod:obl*. Obviously, these annotations had to be carried out manually as most of these cases could not be easily and unequivocally converted to the UD format only on the basis of morphology and syntax. Consider the following examples (Example 19 is repeated for convenience):

(22) Nekem nehogy eladd
I-DAT so.as.not.to sell-IMP-2SGOBJ
az autódat !
the car-2SGPOSS-ACC !
As for me, you should not sell your car.

(23) Nehogy eladd nekem az
so.as.not.to sell-IMP-2SGOBJ I-DAT the
autódat !
car-2SGPOSS-ACC !
You should not sell your car to me.

Example 22 contains a dativus ethicus whereas Example 23 contains an indirect object. The two sentences only have different word orders thus their automatic distinction would not be straightforward.

4.5 Light verb constructions

Light verb constructions are verb + noun combinations where most of the semantic content of the whole expression is carried by the noun while the syntactic head is the verb (e.g. *to have a shower*, *to make a decision*). They are not uniformly treated in Version 1.3 of the UD treebanks. Light verb constructions are either not marked at all or if they are marked, they may have a special structure or special labels (Nivre and Vincze, 2015). The Hungarian treebank belongs to the latter group, that is, members of light verb constructions bear a special

name in Hungarian.

label. For instance, Figure 3 shows that the label `dobj:lvc` can be found between the nominal and verbal component of the light verb construction *döntést hoz* (decision-ACC bring) "to make a decision". In this way, the `dobj` part of the label marks that syntactically it is a verb–object relation but semantically, it is a light verb construction, marked by the `lvc` extension of the label.

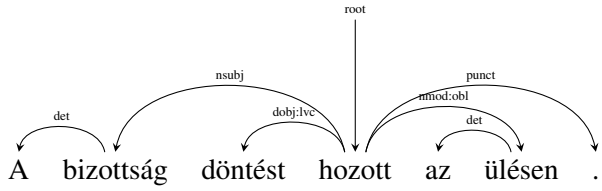

Figure 3: Light verb construction in the Hungarian UD treebank (*A bizottság döntést hozott az ülésen* (the committee decision-ACC bring-PAST-3SG the meeting-SUP) "The committee made a decision at the meeting").

5 Experiments

We developed a converter from the existing Szeged Dependency Treebank (Vincze et al., 2010) to UD and manually corrected 1,800 sentences from the newspaper domain. The manually corrected UD sentences are available in the UD repository v3.0. The experiences gained during the manual correction could reinforce the linguistic conversion rules and the manually corrected gold standard corpus provides the opportunity for empirical evaluations which we introduce in this section.

5.1 On the Accuracy of Automatic Converters

Most of the UD treebanks are the result of automatic conversion from a dependency treebank of originally different principles. The accuracy of these automatic converters is unknown, i.e. we do not know how much information was lost or how much noise was introduced by the converters. To empirically investigate this in the case of Hungarian UD, we compared the converted and the manually corrected, i.e. gold standard, trees of the 1800 sentences.

The converter itself is based on linguistic rules (it is available on our website[5]) which were iteratively improved by manually investigating the results of conversion on sentences of the Szeged Dependency Treebank. The final version of the converted achieves an UAS of 87.81 and a LAS of 75.99 on the 1800 sentences compared against the manually corrected UD trees. We believe that this level of accuracy is not sufficient for releasing the rest of the 80,000 sentences of the automatically converted Szeged Dependency Treebank. On the other hand, some of the shortcomings of the automatic conversion could be corrected by exploiting annotation found in other versions of the Szeged Treebank. For instance, the type of certain subordinate clauses is marked in the constituency version of the treebank, which can be transformed into UD labels. Moreover, coreference annotations from the subcorpora annotated for coreference relations could enhance the proper attachment of relative clauses. We intend to add these pieces of information to our converter in the future, hence higher accuracy scores can be provisioned for the automatic conversion process: just with the above mentioned corrections, an additional 6 percentage points could be achieved in terms of LAS as about 20% of the errors are due to subordinate or relative clauses.

[5] `http://rgai.inf.u-szeged.hu/dependency`

5.2 On the Price of Universality

We carried out experiments for investigating whether is there any difference between using the original MSD (Vincze et al., 2014) and the new universal morphological (UM) descriptions. We were particularly interested in the utility of the two representations for dependency parsing. We trained two models of the MarMot morphological tagger (Mueller et al., 2013) using the two morphological representation in 10-fold cross-tagging on our manually corrected 1800 sentences. Then we trained and evaluated the Bohnet dependency parser (Bohnet, 2010) on the train/test split of the UD repository v3.0 utilizing the two different predicted morphological descriptions. We used the default parameters for both the MarMot and the Bohnet parser.

Table 1 presents unlabeled (UAS) and labeled (LAS) attachment scores achieved by the parser on the test set. The first column of the table indicates whether the universal morphology (UM) or the original MSD morphological codes were employed in the experiment. The second column of the table shows which dependency label set

Morph. labels	Dep labels	UAS	LAS (full label)	LAS (main label)
UM	full label	81.94	76.98	78.39
MSD	full label	82.27	77.50	78.75
UM	main label	81.70	–	78.39
MSD	main label	82.17	–	78.58

Table 1: Dependency parsing results on the Hungarian Universal Dependency dataset. In the case of $LAS(main\ label)$ we do not check the language specific part of the dependency labels in the evaluations while we compare the universal and language-specific dependency labels at $LAS(full\ label)$.

was used for training the Bohnet parser. *main label* refers here to the universal dependency labels while *full label* refers to using the concatenation of universal and language-specific labels. The difference between the last two columns of the table is that we checked the full or only the main dependency labels at evaluations.

Table 1 shows the MSD outperforms UM consistently at each of the experiments. Although these differences are not high, this suggests that some information encoded in the MSD morphology is not represented in UM, i.e. we have to pay a price to be universal. We can observe the greatest difference when training and evaluating on full dependency labels, i.e. language-specific morphological features contribute to the prediction of language-specific dependency labels.

We made a manual error analysis of the results with regard to attachment (UAS) errors, i.e. we compared the outputs of the dependency parsers trained by using predicted universal codes and predicted MSD morphological codes, respectively. Results are presented in Table 2. We found that the benefits of the original language-specific annotation (MSD) mostly manifests in the treatment of subordinate clauses, adverbial modifiers and infinitival complements. These results might be explained by the fact that in certain cases, MSD contains more detailed grammatical information than the UM formalism. For instance, MSD encodes whether a conjunction connects clauses or words/phrases, which information is missing from UM. Also, higher results were achieved for cases when two nouns or adjectives were following each other and one of them modified the other (as in *magas rangú képviselői* "representatives of high standings"). However, sentences containing an overt or covert form of the copula could be parsed more effectively by using universal morphology codes.

Error type	MSD	%	UM	%
Coordination	100	32.05	98	30.34
Article	44	14.10	44	13.62
Adverbial	35	11.22	43	13.31
Other	37	11.86	31	9.60
Part/adj compl.	31	9.94	32	9.91
Adjacent N/A	15	4.81	20	6.19
Subordination	13	4.17	17	5.26
Copula	14	4.49	11	3.41
Infinitive	9	2.88	15	4.64
Nominal arg.	8	2.56	8	2.48
Possessor	6	1.92	4	1.24
Total	312	100	323	100

Table 2: Error analysis: the number and ratio of specific error types.

5.3 The Added Value of Language-specific UD Labels

We also investigated the impact of the language-specific parts of the dependency labels. As the numbers in Table 1 show, slightly better results can be achieved both in terms of UAS and LAS when training the model with *full label*s than with *main label*s. This highlights the importance of adding language specific distinctions to the universal ones because they may contain information that can be exploited during the tree decoding. They contribute even to unlabeled attachment decisions. To take an example, UD does not make any distinction among different types of nominal modifiers, treating them as `nmod`. However, for Hungarian, we applied extra labels such as `nmod:poss` for possessors (see Section 3.1) and `nmod:obl` for nominal arguments of the verb. As for the first, it should always be attached to the possessed noun, whereas the second one is attached to a verb (see also Examples 18 and 19 with the dative morphological case). Thus, the parser can learn these fine-grained distinctions, which might be beneficial for the unlabeled attachment scores as well.

Also, we would like to point out that the utilization of language-specific labels does not contradict the UD principles. In UD, each language should select the appropriate labels according to their needs but there is no need to apply all of the labels/features. General labels like `nsubj` or `dobj` will be used in most (maybe all) of the UD languages but there are other labels or feature-value pairs that are applicable for only a handful of languages. These ones are now called as "language-specific" features but in principle, their status is not different from those that are more widely applied. So we believe that introducing "language-specific" additions does not harm the UD principles. Moreover, the chief objective of our experiments was to highlight the added value of language-specific features and we were able to show that they can even improve parsing accuracy when evaluated exclusively on the general labels. The main goal of UD is to provide a way where the parsing results over languages are comparable, hence using language specific features during decoding but evaluating only on general labels is in line with this comparison principle. Moreover, it indicates for UD treebank developers that – besides general labels – language-specific ones have to be taken seriously.

6 Conclusions

In this paper, the principles of universal dependencies and morphology for Hungarian were introduced by reporting the most challenging grammatical phenomena and our solutions to those. We converted then manually corrected 1,800 sentences from the Szeged Treebank to universal dependency format and introduced experiments on this manually annotated corpus for evaluating automatic conversion and the added value of language-specific, i.e. non-universal, annotations. We would like to draw the attention to the importance of understanding i) the information loss of the automatic UD converters; ii) what is the price of being constrained by universal morphology principles and; iii) the utility of exploiting language-specific dependency labels in UD.

Acknowledgments

The research of Richárd Farkas was funded by the János Bolyai Scholarship.

References

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Illarraza, Iakes Goenaga, Koldo Gojenola, and Larraitz Uria. 2014. Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies. In *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14)*.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97.

Dóra Csendes, János Csirik, and Tibor Gyimóthy. 2004. The Szeged Corpus. A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In *COLING 2004 5th International Workshop on Linguistically Interpreted Corpora*, pages 19–22.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

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

Alexander Fraser, Helmut Schmid, Richárd Farkas, Renjing Wang, and Hinrich Schütze. 2013. Knowledge sources for constituent parsing of german, a morphologically rich and less-configurational language. *Computational Linguistics*, 39(1):57–85.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for finnish: the turku dependency treebank. *Language Resources and Evaluation*, 48(3):493–531.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2014. Universal Dependencies for Danish. In *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14)*.

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.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332.

Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. 2016. Estonian Dependency Treebank: from Constraint Grammar tagset to Universal Dependencies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Joakim Nivre and Veronika Vincze. 2015. Light verb constructions in universal dependencies. Poster at the 5th PARSEME meeting, Iaşi, Romania.

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, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Joakim Nivre. 2014. Universal Dependencies for Swedish. In *Proceedings of SLTC 2014*.

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

Lilja Øvrelid and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

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

Alain Polguère and Igor Aleksandrovič Mel'čuk, editors. 2009. *Dependency in Linguistic Description*. Studies in language companion series.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *Proceedings of Nodalida*.

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*.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Yuval Marton, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, and Alina Wróblewska. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 Shared Task on Parsing Morphologically-rich Languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109.

Mojgan Seraji, Filip Ginter, and Joakim Nivre. 2016. Universal Dependencies for Persian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal Dependencies for Japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

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

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian Dependency Treebank. In *Proceedings of LREC 2010*.

Veronika Vincze, Viktor Varga, Katalin Ilona Simkó, János Zsibrita, Ágoston Nagy, Richárd Farkas, and János Csirik. 2014. Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. In *Proceedings of LREC 2014*, pages 1074–1078.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Daniel Zeman. 2015. Slavic Languages in Universal Dependencies. In *Slovko 2015: Natural Language Processing, Corpus Linguistics, E-learning*.