Variety delights (sometimes) – Annotation differences in morphologically annotated corpora

Andrea Dömötör[1], Balázs Indig[1,2], Dávid Márk Nemeskey[1],
[1]National Laboratory for Digital Heritage
[2]ELTE Faculty of Informatics
{domotor.andrea,nemeskey.david}@btk.elte.hu
indig.balazs@inf.elte.hu

Abstract

The goal of annotation standards is to ensure consistency across different corpora and languages. But do they succeed? In our paper, we experiment with morphologically annotated Hungarian corpora of different sizes (ELTE DH gold standard corpus, NYTK-NerKor, and Szeged Treebank) to assess their compatibility as a combined training corpus for morphological analysis and disambiguation. Our results show that combining any two corpora not only failed to improve the results of the trained tagger, but even degraded them due to the inconsistent annotations. Further analysis of the annotation differences among the corpora revealed inconsistencies of several sources: a different theoretical approach, lack of consensus, and tagset conversion issues.

Keywords: morphology, corpus annotation, corpus evaluation, POS tagging

1 Introduction

Annotation standards such as Universal Dependencies (UD) (Nivre et al., 2017) are intended to facilitate consistent annotation across corpora and languages. Linguistic annotation is time-consuming; therefore, combining different corpora that share the same annotation scheme could be an effective strategy to increase corpus size. In our research, we explored this possibility with morphologically annotated corpora in Hungarian. Training a text processing tool with several different Hungarian corpora has previously been proven to be an effective method for the recognition of named entities (Simon et al., 2022). Our assumption was that a larger training corpus would increase the performance of a lemmatizer and morphological analyzer tool as well.

However, linguistic annotation is a complex task and different theoretical approaches may allow subjectivity even within a well-defined annotation scheme. Therefore, it is highly questionable whether the corpora that are expected to be compatible are indeed so; and if not, whether it is possible to ensure a higher level of compatibility without manually re-annotating one of them.

In this paper we examine the compatibility of three morphologically annotated Hungarian corpora by using them as training data for POS-tagging tools. In Section 3 we present the corpora, their tagsets, and the tagger tools in detail. The section also describes our experiment setup: each corpora was split into train, dev, and test subsets which we used in different combinations for training and testing. Our results presented in Section 4 showed that pairing different corpora lowered the performance in each case. To analyze the differences in the tagsets and annotation schemes of the corpora, we performed further training and testing experiments where we used one corpus for training and another for testing (Section 5). The error analysis of these revealed inconsistencies of several sources: a different theoretical approach, lack of consensus, and tagset conversion issues.

Our findings contribute to the standardization of annotation schemes for Hungarian, including the revision of the UD guidelines. We also detected some issues in the corpora and the UD-conversion tool that we used that need to be addressed in the future.

2 Related Work

The issue of combining different corpora was previously addressed by Straka and Straková (2017) in the evaluation of UDPipe version 1.1. They trained the pipeline on a wide range of languages where multiple UD corpora were available. The tagger and parser models were trained both on the individual corpora and on combinations of different corpora. Generally, they found that the models achieved better results when only one corpus was used for training, combining different corpora de-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 269–277
July 5–10, 2020. ©2020 Association for Computational Linguistics

graded performance. They also conducted more detailed experiments for smaller corpora with the goal of examining the possibility to enrich limited training data from other corpora. The paper shows the results in those cases only where the enrichment of the training corpus resulted in better performance in dependency annotation. This means a total of 12 corpora in ancient Greek, Czech, English, French, Italian, Latin, Slovenian, and Swedish languages. Extending the original datasets from other corpora improved the performance of POS tagging in 6 cases, morphological feature identification in 4, and lemmatization in 7 cases. Thus, increasing corpus size from other sources did not work in every case, not even for small corpora. The authors explain this with the inconsistencies in the annotations of the different corpora (*"the Universal Dependencies are yet not so universal as everyone would like"*).

Wisniewski and Yvon (2019) examine the discrepancies in annotations of UD corpora, focusing primarily on English and French treebanks, as these are among the most extensively represented languages. To detect differences between the corpora, they used the method of Boyd et al. (2008), which states that if two identical sequences are annotated differently, then one of the sequences is likely to be inconsistent. According to Wisniewski and Yvon (2019), inconsistencies may naturally occur within a corpus as well, but in all the cases examined, the ratio of conflicting annotations was higher between different corpora than within one. The authors conducted another experiment to characterize differences between corpora. In this, they trained a binary classifier to decide which of the two corpora a sentence belongs to. The intuitive assumption is that the higher the error rate of this classifier is, the more similar the two corpora are. The classifier was trained on words, POS tags, and word + POS tag pairs. The most successful classification was achieved with the last combination, which suggests that varying annotations of identical words (or sequences of words) characterize the corpora well, indicating that the differences between the annotations of different corpora are systematic.

It can thus be said that the discrepancies in annotation schemes among different corpora of the same language are a known issue that affects multiple languages.

3 Corpora and Tools Used

For our experiments, we used three manually annotated Hungarian corpora of different sizes. The largest among them is the Szeged Treebank (Vincze et al., 2010), which is currently used as the training corpus for HuSpacy (Orosz et al., 2023). Its total size is 1 362 505 tokens. The bulk of the original annotations (Csendes et al., 2004) was automatically converted to the Universal Dependencies standard[1]. On a small part of the corpus[2] (42 032 tokens), the converted UD annotations were manually checked and corrected; this is the only subset openly available in the UD treebank repository (Nivre et al., 2020).

The second largest corpus we used is NYTK-NerKor[3] (Simon and Vadász, 2021), which contains a total of 1 017 340 tokens, while the smallest ELTE DH gold standard corpus (K. Molnár and Dömötör, 2023)[4] consists of 496 060 tokens. Both corpora were annotated with the same methodology. They used the emtsv (Indig et al., 2019) text processing pipeline for pre-processing, and its output was manually corrected by human annotators. The rule-based morphological analyzer module (Novák et al., 2016) of the pipeline assigns all possible morphological and morphosyntactic analyses to each word of the input text. The annotations are linked to each morpheme of the word (Example 1). The POS tagger module, PurePos (Orosz and Novák, 2013) disambiguates the analyses suggested by the analyzer module and provides the lemma and the morphological tag of the word (Example 2). The emtsv tag is a simplified combination of the emMorph tags of each morpheme of the word.

(1) *tető[/N]-n[Supe]*
roof-SUPESSS

’on (the) roof’

(2) Word: *tetőn – ’on (the) roof’*
Lemma: *tető – ’roof’*
Tag: `[/N][Supe]`

[1] https://github.com/huspacy/huspacy-resources/tree/master/data/processed/szeged-corpus

[2] https://github.com/UniversalDependencies/UD_Hungarian-Szeged/

[3] https://github.com/nytud/NYTK-NerKor/

[4] https://github.com/ELTE-DH/gold-standard

This means that the emtsv tags are not merely POS-tags. They also contain all the morphosyntactic information that is represented in the morphological features in Universal Dependencies. The emtsv tagset can be converted automatically to UD; both NerKor and the ELTE DH corpus used the emmorph2ud2 (Vadász and Simon, 2019) converting tool to add the UD annotation layer. The UD tags were not manually checked in either of the corpora, but NerKor did apply some dictionary- and rule-based corrections in cases where their scheme differed from the UD guidelines[5]. The ELTE DH corpus did not change the output of the UD conversion tool (as it is supposed to be unambigous).

In summary, all three corpora have UD morphological annotations and two of them also contain emtsv tags, meaning the three corpora could potentially be merged to form a substantially larger and more comprehensive training dataset for morphological analyzers and POS-tagging tools. All three corpora are genre heterogeneous, containing overlapping and unique text types. Combining the corpora thus achieves not only a larger size but also greater genre diversity. The genres found in the corpora are summarized in Table 1.

For testing the compatibility of the corpora, we trained the lemmatizer and morphological analyzer modules of HuSpaCy and PurePos on each. HuSpaCy is a project that provides Hungarian models for spaCy, the latter of which does not officially support the language. Similarly to spaCy, it uses UD POS tags and morphological features. PurePos is an HMM-based automatic morphological annotation tool optimized for the emtsv tagset with the option of pre-analysis using the rule-based emMorph (Novák et al., 2016) module.

For the train-dev-test split of the corpora, we used the division of HuSpaCy's original training data (derived from the Szeged Treebank). The cutting ensured that each subcorpus is represented in the train, dev, and test sets with the same proportion, and that each set contained complete sentences only. First, the corpora were used separately for training and testing, then we attempted to combine them in pairs.

All models were trained for at most 50 epochs. For HuSpaCy, we disabled all components aside from the senter, tagger, morphologizer and lemmatizer modules. Due to inconsistencies in the HuSpaCy dependencies, we were unable to retrain the transformer-based models and only report results for the `hu_core_news_lg`[6] model. For context, these results can be compared with the numbers achieved by the public spaCy (Honnibal et al., 2020) models for other languages. The results of a total of 82 models in 24 languages are available on the official website.[7] The average performance of the models in POS tagging, morphological features identification, and lemmatization is shown in Table 2.

4 Results

4.1 HuSpaCy

Table 3 shows the results of HuSpaCy trained on different corpora and their combinations. In part-of-speech tagging (POS), NerKor achieved the best result. The performances in lemmatization seem to correspond to the sizes of the individual corpora. In identifying morphological features (Feats), the Szeged Treebank significantly underperformed compared to the other two corpora. However, it can generally be said that all three corpora meet or exceed the average performance of spaCy models in other languages, presented in Table 2.

In the bottom part of the table, we see that combining different corpora degraded the results in almost every case. The results of the smallest corpus (ELTE DH) slightly improved when combined with NerKor. In another instance, we see an improvement is the lemmatization accuracy of the ELTE–Szeged pairing, which surpasses that of the ELTE DH corpus but still stays below the accuracy achieved by the Szeged corpus alone. The worst result was obtained by pairing the two larger corpora, NerKor and the Szeged Treebank. According to these results, ELTE DH and NerKor seem more compatible than any other corpus pair. This might be due to the fact that both used the same converter tool to create their UD layers.

4.2 PurePos

We conducted similar experiments with PurePos on the two corpora containing emtsv annotations (ELTE DH and NerKor). First the analyzer was trained without using the emMorph module, meaning it had to learn the tagset solely from the data without pre-analysis available. Similarly to the

[5] https://github.com/nytud/NYTK-NerKor/blob/main/ud_pos_feats.md

[6] https://huggingface.co/huspacy/hu_core_news_lg

[7] https://spacy.io/models

	ELTE DH	**NYTK-NerKor**	**Szeged Treebank**
Literary	✓	✓	✓
Scientific-popular	(articles) ✓	(wikipedia) ✓	
Blog	✓		
Legal	✓	✓	✓
News		✓	✓
Web		✓	
Student essays			✓
IT-related			✓

Table 1: Genres of the corpora

POS	**Morph**	**Lemma**
0,966	0,944	0,940

Table 2: Average accuracy values of spaCy models in different languages

Corpus	**train**	**dev**	**test**	**POS**	**Lemma**	**Feats**
ELTE DH	485 525	5250	5285	0,982	0,975	0,977
NerKor	997 002	10 167	10 148	0,986	0,982	0,979
Szeged	1 340 639	11 418	10 448	0,983	0,987	0,969
ELTE DH + NerKor	1 482 527	15 417	15 433	0,984	0,977	0,978
ELTE DH + Szeged	1 826 164	16 668	15 733	0,976	0,979	0,954
NerKor + Szeged	2 337 641	21 585	20 596	0,914	0,918	0,897

Table 3: HuSpaCy results trained on different corpora

experiments with HuSpaCy, we trained PurePos separately on each corpus as well as on their combination. The results are shown in Table 4. The UD and emMorph lemmas are presented in separate columns because NerKor assigns two types of lemma to the words: the original (emMorph) lemmas were adjusted to the UD scheme during the UD conversion. Thus, we included both lemma variants in our training experiments.

We can see that the two corpora performed equally in the tagging task despite their different sizes. In lemmatization, the UD lemmas of NerKor proved to be easier to learn than the emMorph lemmas, whereas the two types attained the same accuracy in the ELTE DH corpus (which further was incidentally the same as the results for the emMorph lemmas in NerKor). We find again that combining the two corpora not only failed to improve the results but downright degraded them.

Table 5 presents results from the same training setup but this time we used the built-in emMorph pre-analyzer module so the task of the model trained from the corpora was disambiguation only. For reference, it is worth examining how much of the words are already unambiguous. This was most easily measurable in the xml version of the ELTE DH corpus, as it contains all alternative emtsv analyses. Accordingly, for nearly half (45.7%) of the words both the lemma and the tag are unambiguous. This sets a baseline for (and a lower limit on) the performance of PurePos on this corpus.

Compared to Table 4, the results are mixed. The emMorph pre-analyzer improved both the tagging and lemmatization performance on the ELTE DH corpus significantly; in the latter task, PurePos + emMorph even outperforms HuSpaCy. The comperatively lower results on NerKor suggest that the annotations of NerKor tend to differ from the emtsv pre-analyses.

5 Corpus and tagset differences

The results shown in the previous section suggest significant annotation inconsistencies between the

Corpus	train	test	Tag	Lemma (UD)	Lemma (emMorph)
ELTE DH	485 525	10 535	0,948	0,925	0,925
NYTK-NerKor	997 002	20 315	0,948	0,940	0,925
ELTE DH + NerKor	1 482 527	30 850	0,942	0,923	0,919

Table 4: PurePos results trained on various corpora without emMorph pre-analysis

Corpus	train	test	Tag	Lemma (UD)	Lemma (emMorph)
ELTE DH	485 525	10 535	0,963	0,982	0,982
NYTK-NerKor	997 002	20 315	0,936	0,948	0,954
ELTE + NerKor	1 482 527	30 850	0,942	0,958	0,958

Table 5: PurePos results trained on various corpora with emMorph pre-analysis

examined corpora that might be caused by differences in the tagset or in the use of certain tags. In this section we discuss in detail the inconsistencies we found.

5.1 UD POS tags

The UD POS tagsets are quite consistent in the three corpora, we only found two differences. The first one is marginal: Szeged Treebank uses a special SYM tag for emoticons while the other two corpora tag them as X. The other difference, the usage of the AUX (auxiliary verb) tag is more common and problematic. The ELTE DH corpus does not have AUX tag at all and the Szeged Treebank and NerKor tags different words with it.

In the UD guidelines[8] an auxiliary is described as "a function word that accompanies the lexical verb of a verb phrase and expresses grammatical distinctions not carried by the lexical verb". The guidelines differentiate tense, passive, modal, agreement auxiliaries, and verbal copulas within this category. The Hungarian UD guidelines are quite narrow on the issue, it states that "we consider the verbs “volna”, “fog”, “talál” and “szokott” as AUX in Hungarian". *Volna* and *fog* are tense auxiliaries for the past conditional and future tenses respectively, while *talál* and *szokott* express modality (*’happen to’*) and aspect (*’used to’*). This list seems rather arbitrary and none of the corpora adhere to it.

Szeged Treebank uses the AUX tag for the two tense auxiliaries *volna* and *fog*, as well as for copulas. *Volna* has only one form and is attached to a finite verb (Example 3a). *Fog* has the paradigm for person and number and accompanies an infinitive (Example 3b). Finally, the copula is also conjugated for person and number, but it has present and past tenses as well (Example 3c).

The UD tags in the other two corpora are conversions from the emtsv tagset, which does not have an auxiliary tag itself. As the UD conversion in the ELTE DH corpus was fully automatic, the AUX tag is missing from the corpus altogether. In Nerkor, the auxiliary *volna* is tagged as [/V] (verb with no inflections) which allows their automatic conversion to AUX. However, this was not an option for *fog* and the copula as those have inflections and coincide with other verbs (e.g. *fog* also means "to grasp/hold").

(3) a. *Elmondhattad* *volna*
tell-PST-MOD-SG2 COND

’You could have told (me)’

b. *El* *fogja* *mondani*
PVB FUT-SG3 tell-INF

’He/She will tell’

c. *Ez* *gyors*
this-PRON fast-ADJ
volt
was-COP-SG3-PAST

’It was fast’

The UD guidelines mention modal auxiliaries as well, which is controversial in the Hungarian linguistic tradition (Kalivoda and Prószéky, 2024). They are commonly described as finite verb + infinitive constructions, but they do not form a well-defined category. Therefore, annotating them as

[8] https://universaldependencies.org/guidelines.html

AUX would inevitably require arbitrary decisions about which words to include as modal auxiliary.

In order to detect other systematic differences in the annotation schemes of the three corpora, we conducted further experiments where we used one corpus for training and another one for testing. Table 6 shows the POS-tagging results with HuSpaCy.

	ELTE DH	**NerKor**	**Szeged**
ELTE DH	0,982	0,950	0,930
NerKor	0,944	0,986	0,944
Szeged	0,922	0,937	0,983

Table 6: POS-tagging results across corpora. Each row shows the results of the model trained on the corpus indicated in the first column.

Not surprisingly, using the same corpus for training and testing provides the best result. For more insight on annotation differences, we examined the F-scores by tag. We found that most common tags (NOUN, ADJ, VERB, NUM, DET, PART, SCONJ, PUNCT) show stable results with any training and testing setup. Some tags' scores however, drop significantly when the training and testing data are from different corpora.

This is the case with proper nouns (PROPN) that can be explained with annotation differences and anomalies in the UD conversion. Emtsv does not have a specific tag for proper nouns, so the converter tool converts every uppercased noun to PROPN. This can be problematic with multiword proper names that contain adjectives and other words as well, such as certain institution names. The ELTE DH corpus annotates the elements of these based on their morphology; therefore, the adjectival parts of multiword names are converted to ADJ instead of PROPN. NerKor solves this issue by using 'part of proper name' (caseless noun, i.e. [/N]) tags for each inner token in a named entity. With this approach named entities are handled as a whole, and the morphological features of the inner elements are not displayed. Another approach could be to keep the original emtsv tags of the elements and modify the UD converter accordingly (by including uppercased adjectives).

Another common issue is the distinction of coordinate conjuncts (CCONJ), subordinate conjuncts (SCONJ) and adverbs (ADV). The confusion between CCONJ and SCONJ (which happened when Szeged Treebank was paired with another corpus) is likely due to the UD conversion. Emtsv has only one [/Cnj] tag for both coordinate and subordinate conjuncts. The converter differentiates based on a lexicon that lists 10 elements as subordinate conjuncts. Other conjuncts are converted to CCONJ, often wrongly. The list of subordinate conjuncts needs to be extended with elements such as *mintha* 'like/as if', *hogyha* 'if', *minthogy* 'since/whereas', etc.

The confusion between conjuncts and adverbs (and also pronouns) is quite common, as several lexical items are in fact ambigous. A closer look at these tags in the corpora revealed that Szeged Treebank overuses the ADV tag. There are 10 lemmas that Szeged Treebank exclusively tags as ADV while in NerKor and ELTE DH they are (and should be) tagged as conjuncts, such as *emellett* 'besides', *mialatt* 'while' and *ugyanakkor* 'at the same time'. The dropping F-score of the ADV tag in the Szeged – other corpus pairings is likely due to these erroneous annotations.

5.2 UD features

The feature sets of the corpora also show some differences. Szeged Treebank has some unique features that are not present in the other two corpora. Poss is a boolean feature for possessive pronouns, determiners, or adjectives. Szeged Treebank uses it for possessive pronouns, while ELTE DH and NerKor mark the possessiveness of pronouns with the Number[psed] (possessed object's number) feature. Other feature exclusively used in Szeged Treebank is NumType[sem] that is not mentioned in the UD guidelines but according to Szeged Treebank's data it specifies some semantic categories of numeric lexical items such as time (*7.20*), result (e. g. of a futball match: *2:0*) or quotient (*50:50*). The functions of Type and Cas features in Szeged Treebank are not exactly clear. Type is used for website names and gets values of *w* or *o*. Cas is probably an obsolete version of Case where the case values are coded with numbers. Lastly, Szeged Treebank is not consistent with the name of the reflexive pronoun feature. It appears both in form of Reflex (which is the correct form according to the UD guidelines and is used in the other two corpora) and Reflexive.

There are slight differences in the feature value sets as well. Some values are not represented in all three corpora because they are rare. This is the case with the absolute superlative Degree=Abs and the "general locative" Case=Loc used for the archaic locative of some Hungarian cities. Other value differences are caused by the UD conversion of emtsv. The dative and genitive cases have the same suffix in Hungarian (*-nak/-nek*, see Example 4) and emtsv always annotates them as dative, there is no tag for the genitive case. Therefore, the UD converter converts all nominals with the dative/genitive suffix to dative, which means that the ELTE DH corpus has no `Case=Gen` feature value. NerKor, however, seems to have changed some of the `Case=Dat` values to genitive, probably with the intention of matching Szeged Treebank. The method of identifying the genitive case is not documented thus it is unsure whether the `Case=Gen` features are correct.

(4) a. *a* *cég* *elemző-i-nek*
the company analyst-PL-GEN
közlés-e
announcement-POSS.SG3

’the announcement of the company’s analysts’

b. *átad-t-a* *a* *cég*
hand-PST-SG3 the company
elemző-i-nek
analyst-PL-DAT

’He/She handed it/them to the company’s analysts’

Other difference between ELTE DH and NerKor is that NerKor distinguishes between adjectival participles and adjectives, using `[/V][_ImpfPtcp/Adj]`, `[/V][_PerfPtcp/Adj]`, and `[/V][_ModPtcp/Adj]` tags for the former, while in the ELTE DH corpus, this distinction only appears in detailed emMorph analysis; the simple emtsv tag is `[/Adj]` in every case. While the UD converter converts both adjectives and participles to `ADJ`, the difference still affects the UD features, as in NerKor an extra `VerbForm` feature is added for participles, which does not appear in either the ELTE DH or the Szeged Treebank, where the annotation for adjectival participles matches that of simple adjectives.

Another issue with the UD conversion is that it loses some cases that are present in emtsv. For example, the comitative case is not handled at all by the converter script; therefore, it converts to the default nominative. Nouns in the distributive case are converted to `ADV` which results in dropping all the features. As the derivational suffix for the distributive case is productive, the noun POS tag and the `Case=Dis` feature should be kept.

Lastly, Szeged Treebank has some erroneous `PronType` values, like `PrsPron` instead of `Prs` or pronoun types coded with single letters (probably a remainder from an older version of the corpus).

The overall results of the features with train and test sets of different corpora are shown in Table 7. It seems that ELTE DH and Szeged Treebank make the least compatible pairing. This is probably mostly due to the previously mentioned conversion issues, some of which have been corrected in NerKor.

	ELTE DH	**NerKor**	**Szeged**
ELTE DH	0,977	0,931	0,896
NerKor	0,926	0,979	0,925
Szeged	0,889	0,906	0,969

Table 7: Feature results across corpora. Each row shows the results of the model trained on the corpus indicated in the first column.

Examining the F-scores by feature revealed that pairing different corpora makes the results of `NumType` and `PronType` features drop the most (in addition to those already mentioned). The most confused values of the `NumType` feature are `Card` (cardinal numbers) and `Frac` (fractions). A notable difference we found in the use of these values is that Szeged Treebank uses the `Frac` value for numbers with decimals while these numbers have `NumType=Card` values in ELTE DH and NerKor. The main issue with `PronType` is the distinction of personal (`Prs`) and demonstrative (`Dem`) pronouns, especially between ELTE DH and Nerkor. Emtsv has different tags for these pronoun types (`[/N|Pro]` and `[/Det|Pro]`, respectively) that were often confused by the PurePos models with every corpus setup. After the UD conversion, both pronouns get the `PRON` POS tag; they only differ in the `PronType` feature. Although personal and demonstrative pronouns are often homonymous in Hungarian, the generally low scores of these pronoun types suggest that it might be worth checking their annotations for possible errors.

5.3 emtsv

The emtsv tags of NerKor and ELTE DH are inherently very diverse, as they include several features. According to Vadász and Simon (2019), there are

2088 possible combinations[9]. The two corpora together contain 2024 different tags, only 1025 of which are common between them. This emphasizes the relevance of rule-based analyzer modules (like the emMorph module in PurePos) because a tag variation this great is almost impossible to cover with a training corpus. As emtsv was designed specifically for Hungarian it has several features that are not present in Universal Dependencies. For comparison, the three discussed corpora have altogether 1790 UD POS + feature combinations, 593 of which are common among them. We mapped these UD POS + feature combinations with their respective emtsv tags and found that nominals (nouns, adjectives, and proper nouns) show the greatest diversity. Special features include derivations, semantic categories (like nations or colors), and syntactic (like attributive a predicative adjectives) and word form (like abbreviations and acronyms) features. This much granularity in the tagset is not ideal for machine learning but it can be very valuable for corpus linguists.

The results of PurePos when trained and tested on different corpora are shown in Table 8. As expected, the performance of the models is 4-5% lower in the cross-evaluation setup.

	ELTE DH	**NerKor**
ELTE DH	0,948	0,891
NerKor	0,902	0,942

Table 8: PurePos tagging results across corpora. Each row shows the results of the model trained on the corpus indicated in the first column.

The main differences beetwen the annotation schemes of ELTE DH and NerKor were already discussed in the previous sections. With the UD conversion these differences split between the POS tags and the features.

6 Summary

In summary, the consistency of annotations proved to be more crucial than corpus size in training morphological analyzers. The results obtained from the combination of different corpora demonstrated that even small discrepancies in the annotation schemes can pose significant challenges to the tagging tools.

The annotation differences of the corpora are from several sources. In some cases they are deliberate like the different handling of multiword proper names in ELTE DH and NerKor. Annotations may also differ due to the lack of consensus regarding a phenomenon or category, which is the case with auxiliaries in Hungarian. In other cases the cause of difference was the fact that one of the corpora over-simplified (or complicated) a tag or simply made mistakes. An example for the former is the different annotations of participles in ELTE DH and NerKor, and for the latter we can mention the overuse of `ADV` in Szeged Treebank, mostly at the expense of conjuncts.

Our research also revealed some issues with the emtsv–UD converter tool. For future work we plan to extend the list of subordinate conjuncts and add the missing cases.

As we got good results with training with the corpora separately, the question arises whether compatibility of different corpora is really that essential. In our opinion, having detailed guidelines is crucial for an international standard like Universal Dependencies. The fact that this is still missing for Hungarian presents an ongoing challenge for the Hungarian NLP community. Fixing the issues revealed in our research, such as the obsolete features in Szeged Treebank and the annotation of participles in ELTE DH, is also an important future work.

However, emtsv is an inherently language-specific annotation scheme for Hungarian, which makes the emMorph analysis and the emtsv tag layer a suitable way for the corpora to retain their unique character.

References

Adriane Boyd, Markus Dickinson, and Detmar Meurers. 2008. On detecting errors in dependency treebanks. *Research on Language and Computation*, 6:113–137.

Dóra Csendes, János Csirik, and Tibor Gyimóthy. 2004. The szeged corpus: A pos tagged and syntactically annotated hungarian natural language corpus. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora LINC 2004 at The 20th International Conference on Computational Linguistics COLING 2004*, pages 19–23.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai. 2019. One

[9] https://github.com/nytud/panmorph/blob/master/emmorph.tsv

format to rule them all – the emtsv pipeline for hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop. Association for Computational Linguistics*, pages 155–165.

Emese K. Molnár and Andrea Dömötör. 2023. Gondolatok a gondola-tokról. Morfológiai annotációt javító módszerek tesztelése gold standard korpuszon. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 341–356, Szeged.

Ágnes Kalivoda and Gábor Prószéky. 2024. Hungarian auxiliaries revisited. *Acta Linguistica Academica*, 71:202–218.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Attila Novák, Borbála Siklósi, and Csaba Oravecz. 2016. A New Integrated Open-source Morphological Analyzer for Hungarian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1315—-1322, Portorož.

György Orosz, Gergő Szabó, Péter Berkecz, Zsolt Szántó, and Richárd Farkas. 2023. Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. In *Text, Speech, and Dialogue*, pages 58–69, Cham. Springer Nature Switzerland.

György Orosz and Attila Novák. 2013. PurePos 2.0: a hybrid tool for morphological disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, pages 539–545, Hissar, Bulgaria. INCOMA Ltd. Shoumen.

Eszter Simon and Noémi Vadász. 2021. Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus. In *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings*, volume 12848 of *Lecture Notes in Computer Science*, pages 222–234. Springer.

Eszter Simon, Noémi Vadász, Dániel Lévai, Dávid Márk Nemeskey, György Orosz, and Zsolt Szántó. 2022. Az NYTK-NerKor több szempontú kiértékelése. In *XXVIII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 403–416, Szeged. Szegedi Tudományegyetem TTIK, Informatikai Intézet.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Noémi Vadász and Eszter Simon. 2019. Konverterek magyar morfológiai címkekészletek között. In *XV. Magyar Számítógépes Nyelvészeti Konferencia*, pages 99–111, Szeged. Szegedi Tudományegyetem, Informatikai Intézet.

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

Guillaume Wisniewski and François Yvon. 2019. How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 218–227, Minneapolis, Minnesota. Association for Computational Linguistics.