Szeged Corpus 2.5:
Morphological Modifications in a Manually POS-tagged Hungarian Corpus

Veronika Vincze[1], Viktor Varga[2], Katalin Ilona Simkó[2],
János Zsibrita[2], Ágoston Nagy[2], Richárd Farkas[2], János Csirik[1]

[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
H-6720 Szeged, Tisza Lajos krt. 103.

[2]University of Szeged, Department of Informatics
H-6720 Szeged, Árpád tér 2.

Email: vinczev@inf.u-szeged.hu, viktor.varga.1991@gmail.com, kata.simko@gmail.com,
zsibrita@inf.u-szeged.hu, nagyagoston@inf.u-szeged.hu, rfarkas@inf.u-szeged.hu, csirik@inf.u-szeged.hu

Abstract

The Szeged Corpus is the largest manually annotated database containing the possible morphological analyses and lemmas for each word form. In this work, we present its latest version, Szeged Corpus 2.5, in which the new harmonized morphological coding system of Hungarian has been employed and, on the other hand, the majority of misspelled words have been corrected and tagged with the proper morphological code. New morphological codes are introduced for participles, causative / modal / frequentative verbs, adverbial pronouns and punctuation marks, moreover, the distinction between common and proper nouns is eliminated. We also report some statistical data on the frequency of the new morphological codes. The new version of the corpus made it possible to train *magyarlanc*, a data-driven POS-tagger of Hungarian on a dataset with the new harmonized codes. According to the results, *magyarlanc* is able to achieve a state-of-the-art accuracy score on the 2.5 version as well.

Keywords: corpus, morphology, POS-tagging, Hungarian

1. Introduction

The Szeged Corpus is the largest manually annotated corpus of Hungarian in which all the possible morphological analyses and lemmas for each word form are provided, besides, texts are also POS-tagged (Csendes et al., 2005). In Szeged Corpus 2.0, the MSD morphological coding system is used (Erjavec, 2004). In this work, we present the latest version of the corpus – Szeged Corpus 2.5 – in which we applied some morphological modifications which we believe will benefit real-world NLP applications. The modifications involve the introduction of new codes in the coding system as well as the correction of some morphological codes, with special emphasis on misspelled words.

2. Morphological Coding Systems of Hungarian

There are three widely used morphological coding systems for Hungarian: Humor, MSD and KR. The coding system Humor is based on unification, that is, one word form can contain only morphemes that contain no contradictory morphological features (Prószéky & Tihanyi, 1993).

The MSD morphological coding system is a positional coding system developed for several languages (Erjavec, 2004). By convention, lemmas contain derivational suffixes and only inflectional morphemes are distinguished separately from the lemma.

The KR coding system was developed with respect to Hungarian (Kornai et al., 2004). Linguistic information is encoded in hierarchical attribute value matrices: there are default values (e.g. singular or 3rd person) and only those that differ from these manifest in the code.

3. Harmonizing Morphological Coding Systems of Hungarian

In order to carry out any natural language processing tasks for Hungarian, a basic linguistic preprocessing toolkit is necessary. There are several morphological analyzers available, however, they are based on different morphological coding systems. Thus, a prerequisite for the merge of Hungarian morphological analyzers is the harmonization of the coding systems.

Recently, there has been a successful attempt to harmonize the coding systems MSD and KR (Farkas et al., 2010). It was necessary mostly for the following reasons. *morphdb.hu* is one of the most widely used morphological databases for Hungarian, which is based on the KR morphological annotation system (Trón et al., 2006). However, the Szeged Corpus, the only manually POS-tagged corpus (Csendes et al., 2005) is annotated with MSD codes. The two coding systems cannot be mapped in a one-to-one way, so if we want to exploit both resources in a statistical language parser (POS tagger, constituency parser, dependency parser etc.), we have to employ conversion rules, which leads to the loss of information. In order to prevent this, the two coding systems (MSD and KR) were harmonized and their basic principles were also made compatible. When harmonizing the two coding systems, the following principle was observed: morphological codes should include only those types of information that are useful for later processing

(syntax, applications).

3.1. Derivational suffixes

One of the most important differences is the treatment of derivational suffixes: KR works with absolute stems while MSD works with relative stems, that is, lemmas include derivational suffixes in the latter case and it is only inflectional suffixes that are cut off the word forms. In this case, we adapted from both coding systems those distinctions that can be justified from a higher-level point of view for NLP applications. Thus, manually annotating absolute lemmas / stems would have been an enormous task, moreover, relative lemmas usually provide enough information for applications like information extraction or retrieval, so the harmonized coding system applies relative lemmas.

However, in certain cases, it was necessary to diverge from the above convention. For instance, in the case of derived verbs, only those pieces of derivational information are explicitly marked that are expressed with syntactic tools in other languages. For instance, *olvasgathatják* (read-FREQ-MODAL-3PL.OBJ) 'they can frequently read it', where the lemma is *olvas* 'read', the derivational suffixes *-gat* and *-hat* denote frequentative aspect and modality, respectively, and the morphological code of the word form includes information on frequentative aspect and modality as well. However, no derivational information is marked in the case of the deadjectival verb *szépít* 'beautify', which is derived from *szép* 'beautiful', since this information is irrelevant from a syntactic point of view. We applied the same approach to verbs with frequentative, modal and causative suffixes and the lemma became the word form without any of the above mentioned suffixes.

The second position of the verbal MSD codes represents information on the verb type and the lemma of the verb is the base form. We also paid attention to the fact that these suffixes are not mutually exclusive, that is, a given verb form may be modal and causative at the same time for instance. Hence, all the combinatorial possibilities are listed among the possible codes within the harmonized coding system. Table 1 shows the verbal codes.

We annotated each word form with the new morphological codes, and whenever the word form was ambiguous among several morphological analyses, we manually chose the correct one according to the context. Such cases needed special attention: for instance, in the past tense, the causative and non-causative forms of the same verb may coincide: *festetted* may mean 'you painted it' (paint-PAST-2SG.OBJ) or 'you made someone paint it' (paint-CAUS-PAST-2SG.OBJ), depending on the context.

Verb type	**Code**	**Suffix**	**Example**
main	m	-	*megy* 'go'
auxiliary	a	-	*fogok (menni)* 'I will (go)'
modal	o	-hAt	*mehetek* 'I can go'
frequentative	f	-gAt	*pofozgat* 'he is slapping something'
causative	s	-(t)At	*etet* 'feed' (lit. 'make eat')
frequentative + modal	1	-gAthAt	*boncolgathat* 'he can be analyzing something'
causative + modal	2	-(t)AthAt	*fektethet* 'he can lay down something' (lit. 'he can make something lay down')
causative + frequentative	3	-(t)AtgAt	*etetget* 'he is feeding' (lit. 'he makes someone eat frequently')
causative + frequentative + modal	4	-(t)AtgAthAt	*futtatgathat* 'he can run something (on a computer)' (lit. 'he can make something run frequently')

Table 1: Verbal harmonized codes.

3.2. Participles

Present, past and future participles were also given a new code since in the earlier version of the corpus, they could not be distinguished on the basis of their codes, what is more, their code coincided with that of adjectives. However, normal adjectives and participles exhibit different grammatical features (for instance, participles cannot be used in comparative/superlative forms), which may be relevant for syntax and thus, this distinction is again justifiable.

The second position of the adjectival MSD code denotes whether it is an adjective or a participle. In the latter case, it is also encoded whether it is a past / present / future participle. Codes are listed in Table 2.

Type	Code	Suffix	Example
adjective	f	-	*friss* 'fresh'
present participle	p	-Ó	*sétáló* 'walking'
past participle	s	-t/-tt	*megvásárolt* '(something) bought'
future participle	u	-AndÓ	*felveendő* '(something) to be recorded'

Table 2: Adjectival and participial harmonized codes.

Some word forms may be used as adjectives and participles as well, e.g. *égető kérdések* 'burning questions'– *a kertben tüzet égető gondnok* 'the housekeeper burning a fire in the garden'. We applied linguistic tests to distinguish between the participial and adjectival uses of the same word when manually annotating the data.

3.3. Common nouns and proper nouns

We also eliminated the differentiation between proper nouns and common nouns at the level of morphology since we believe that it is the task of a named entity recognition system to identify named entities (proper names) in texts rather than that of a morphological parser. Thus, the morphological code of each noun starts with Nn- now.

3.4. Adverbial pronouns

The treatment of adverbial pronouns was one of the most dubious questions of harmonization. In MSD, word forms like *mögötted* (behind-2SG.POSS) 'behind you' or *velünk* (instrumental.suffix+1PL.POSS) 'with us' were coded as subtypes of adverbs, marking only its number and person. On the other hand, they were coded as nouns in KR: the lemma of those derived from a case suffix such as *velünk* was the personal pronoun (in this case, *mi* 'we') and its case was assigned similar to nouns. As for those derived from postpositions such as *mögötted*, the code itself contained the original postpositions, for instance, *mögötted* as coded as te/NOUN<POSTP<MÖGÖTT>>, or Rl--s1 (*mögött*).

In this case, we did not apply any of the previously developed solutions but we argued for deriving all these forms from personal pronouns and thus inserted them into the pronominal system of morphological codes. Table 3 offers some examples of the new annotation scheme.

Word form	Lemma	Morphological code
szerintem 'according to me'	*szerint* 'according to'	Pp1-sn
nálunk 'at us'	*mi* 'we'	Pp1-p3

Table 3: Harmonized codes for adverbial pronouns.

These words were automatically relabeled in the corpus, and no further manual checking was required.

3.5. Punctuation marks

The morphological coding of punctuation marks was also changed. Eight punctuation marks were considered as relevant (they are followed by their ASCII code): !(33) ,(44) -(45) .(46) :(58) ;(59) ?(63) –(8211).

The lemma and morphological code of the relevant punctuation marks are the punctuation mark itself in the harmonized version of the corpus. For other non-relevant punctuation marks (i.e. character strings that do not contain letters or digits), the lemma is the punctuation mark itself but the morphological code is K.

3.6. Separable verbal prefixes

Verbal prefixes in Hungarian may occur right before the verb, in which case they are spelt as one word. In other cases, they can be separated by other words within the sentence or the prefix can follow the verb. In such cases, they are spelt as two words.

In the verbal elements (verbs, infinitives, participles) that contain a separable verbal prefix, the morpheme boundary between the prefix and the verbal element was distinctively marked. Since there are some syntactic processes that trigger the separation of the two elements, we encoded this boundary in the lemma of the given word.

4. Correcting Misspelled Words

In addition to the morphological modifications described above, we also paid attention to the correction of misspelled words. In the 2.0 version of the corpus, misspelled words had a separate morphological code (e.g. *kiráj* instead of the standard spelling *király* 'king / cool' – the combination *ly* denotes the same sound as *j* in Hungarian). So did words that are possible word forms but in the current context, they are improperly applied. For instance, the standard form of the phrase *mer úgy gondolom* (dare so think-1SG.OBJ) would be *mert úgy gondolom* 'because I think so': *mer* is an existing Hungarian verb meaning 'dare' but its usage is improper in this context, thus its morphological code indicates this anomaly.

When the correct and the misspelled forms both contained the same amount of tokens (e.g. *aszt – azt* 'that one-ACC'), the misspelled words were added their correct forms together with their possible morphological analyses and lemmas, and later on, the actual analysis was manually selected according to the context. When the number of tokens differed in the case of the correct and misspelled words, the code of the main element was added to the misspelled unit as in *areggel* (the.morning) vs. *a reggel* (the morning), where the one-token unit *areggel* was tagged as a noun.

5. Statistical Data

Szeged Corpus 2.5 contains 82,000 sentences and 1.2

million tokens. In version 2.0, the number of unknown or misspelled words was 11,461, which number was reduced to 1,563 in version 2.5. Thus, the proportion of unknown or misspelled words (which might be problematic for morphological parsing) changed from about 1% to 0.13%, which means a considerable reduction of erroneous words (86.4% of them were eliminated, in other words, there is a difference of an order of magnitude). Now most of the unknown words are foreign (especially English) terms as in the computer texts subcorpus, user manuals often include the English terminology as well.

In Szeged Corpus 2.5, there are 1315 morphological codes altogether. Table 4 represents the occurrences of the newly introduced codes:

Type	**Code**	**Frequency**
Present participle	Ap*	23,483
Past participle	As*	12,588
Future participle	Au*	520
Participles - total	**Ap*, As*, Au***	**36,591**
Causative verb	Vs*	1,698
Modal verb	Vo*	8,415
Frequentative verb	Vf*	327
Combination of causative / modal / frequentative	V1*, V2*, V3*, V4*	67
Causative / modal / frequentative - total	**Vs*, Vo*, Vf*, V1*, V2*, V3*, V4***	**10,057**

Table 4: Frequency of new codes.

The reannotation process of adverbial pronouns affected another set of codes, namely, 8232 tokens were reannotated. Thus, if all the words with a new morphological code are counted (participles, causative / modal / frequentative verbs, adverbial pronouns, corrected misspelled words), then we get 64,788 words, which means that about 4% of the words in the 2.5 corpus were reannotated, compared to the previous 2.0 version. This change in morphological data may be fruitfully applied in morphological parsing and POS-tagging.

6. Part-of-speech Tagging

Szeged Corpus 2.5 also made it possible to train *magyarlanc*, a data-driven linguistic preprocessing toolkit of Hungarian (Zsibrita et al., 2013) on the new database. Thus, the morphological analyzer and POS-tagger modules of the toolkit were trained and evaluated on the corpus, which yields a linguistic output that makes use of the new harmonized morphological coding system.

Sentences of the corpus were randomly split into a training and evaluation database in an 80:20 ratio, and we trained and evaluated the POS-tagger module of *magyarlanc* in this way. The analysis provided by *magyarlanc* was considered correct if both the lemma and the morphological code matched with the gold standard data. According to the results, the POS-tagger module of *magyarlanc* achieved an accuracy of 96.32% on Szeged Corpus 2.5 with the new harmonized codes, which coincides with our results published earlier on Szeged Corpus 2.0 (Zsibrita et al., 2013). Thus, it should be noted that the accuracy of POS-tagging does not change significantly when the tagger is trained on a dataset with a larger set of possible morphological codes.

7. Summary

In this work, we presented the 2.5 version of the Szeged Corpus, the biggest manually annotated Hungarian corpus. We described those innovations that have been carried out in the morphological analysis of the word forms, we discussed the treatment of misspelled words and reported results on POS-tagging on the new version. Szeged Corpus 2.5 is freely available for research and educational purposes at the corpus website http://www.inf.u-szeged.hu/rgai/SzegedTreebank.

We hope that this newly annotated corpus will enhance NLP research on Hungarian, especially on morphological and syntactic parsing and furthermore, in the morphological processing of other morphologically rich languages.

8. Acknowledgements

This work was supported in part by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11 / 1 / KONV-2012-0013).

9. References

Csendes, D.; Csirik, J.; Gyimóthy, T.; Kocsor, A. (2005). The Szeged Treebank. In: *Proceedings of the Eighth International Conference on Text, Speech and Dialogue* (TSD 2005). Karlovy Vary, Czech Republic 12-16 September, and LNAI series Vol. 3658, pp. 123--131.

Erjavec, T. (ed.) (2004). *MULTEXT-East morphosyntactic specifications*. Version 3 http://nl.ijs.si/ME/V3/msd/msd.pdf

Farkas, R.; Szeredi, D.; Varga, D.; Vincze, V. (2010). MSD-KR harmonizáció a Szeged Treebank 2.5-ben. In: *VII. Magyar Számítógépes Nyelvészeti Konferencia,* pp. 349--353.

Kornai, A.; Rebrus, P.; Vajda, P.; Halácsy, P.; Rung, A.; Trón, V. (2004). Általános célú morfológiai elemző kimeneti formalizmusa. In: *II. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Szeged, pp. 172--176.

Prószéky, G.; Tihanyi, L. (1993). Humor: High-Speed Unification Morphology and Its Applications for Agglutinative Languages. *La tribune des industries de la langue* 10. OFIL, Paris, France pp. 28--29.

Trón, V.; Halácsy, P.; Rebrus P.; Rung, A.; Simon, E.,

Vajda, P. (2006). Morphdb.hu: Hungarian lexical database and morphological grammar. In: *Proceedings of 5th International Conference on Language Resources and Evaluation* (LREC ’06).

Zsibrita, J.; Vincze, V.; Farkas, R. (2013). magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: *Proceedings of RANLP-2013*, Hissar, Bulgaria, pp. 763-771.