Attila Novák[1,2], Katalin Gugán[3], Mónika Varga[3], Adrienne Dömötör[3]

Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence

[1]MTA-PPKE Hungarian Language Technology Research Group,
[2]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
1083 Budapest, Práter u. 50/a, Hungary

[3]Research Institute for Linguistics of the Hungarian Academy of Sciences
1068 Budapest, Benczúr u. 33., Hungary

Corresponding author: novak.attila@itk.ppke.hu

Abstract

The paper introduces a novel annotated corpus of Old and Middle Hungarian (16–18. century), the texts of which were selected in order to approximate the vernacular of the given historical periods as closely as possible. The corpus consists of testimonies of witnesses in trials and samples of private correspondence. The texts are not only analyzed morphologically, but each file contains metadata that would also facilitate sociolinguistic research.

The texts were segmented into clauses, manually normalized and morphosyntactically annotated using an annotation system consisting of the PurePos PoS tagger and the Hungarian morphological analyzer HuMor originally developed for Modern Hungarian but adapted to analyze Old and Middle Hungarian morphological constructions. The automatically disambiguated morphological annotation was manually checked and corrected using an easy-to-use web-based manual disambiguation interface. The normalization process and the manual validation of the annotation required extensive teamwork and provided continuous feedback for the refinement of the computational morphology and iterative retraining of the statistical models of the tagger.

The paper discusses some of the typical problems that occurred during the normalization procedure and their tentative solutions. Besides, we also describe the automatic annotation tools, the process of semi-automatic disambiguation, and the query interface, a special function of which also makes correction of the annotation possible. Displaying the original, the normalized and the parsed versions of the selected texts, the beta version of the first fully normalized and annotated historical corpus of Hungarian is freely accessible at the address http://tmk.nytud.hu/.

Keywords: Historical corpus, corpus annotation, morphological analysis, PoS tagging, Middle Hungarian, Old Hungarian, corpus query tool

Acknowledgements

The project *Morphologically analysed corpus of Old and Middle Hungarian texts representative of informal language use* was funded by the Hungarian Scientific Research Fund (OTKA) project grant No. OTKA 81189. The participants of the project mainly include historical linguists working at the Department of Finno-Ugric and Historical Linguistics of the Research Institute for Linguistics of the Hungarian Academy of Sciences, but the funding of OTKA made it possible to involve MA and doctoral students as well as participation of the computational linguist of the team. The project greatly benefited from regular consultations with experts of etymology (László Horváth) and historical syntax (Lea Haader). The follow-up project *Competing structures in the Middle Hungarian vernacular: a variationist approach* has been funded by the Hungarian Scientific Research Fund project grant No. OTKA K 116217.

1. Introduction

The main objective of the project described in this article was to facilitate research focusing mainly on the Old and Middle Hungarian vernacular, that is, to compile an electronic collection of types of texts that are assumed to reflect spoken language as closely as possible. The texts that were included in the corpus are court records (trial minutes) and personal letters. In order to make structured linguistic queries on the corpus possible, the texts were annotated with disambiguated morphological analysis. This was accomplished utilizing the Humor morphological analyzer for Modern Standard Hungarian (Novák 2003; Prószéky and Novák 2005)), which was extended to be capable of analyzing words containing morphological constructions, suffix allomorphs, suffix morphemes, paradigms or stems that were used in Old and Middle Hungarian but no longer exist in present-day Hungarian.

For some other, mostly configurational, languages, analyzed historical corpora also contain syntactic annotation. However, for reasons explained in Section 7.2, we decided not to go beyond the level of morphosyntax when defining the objectives of our annotation program. Thus, in our corpus, the texts were first digitized using OCR techniques and manual postcorrection, and then these digital texts were segmented into clauses, and they were normalized. Finally, the texts were analyzed using the adapted morphological analyzer, and they were disambiguated. The disambiguation process was semi-automatic, the analyzer assigning a list of possible analyses to each word, of which a statistical HMM tagger selected a most probable analysis. Finally, the annotation was manually checked and corrected. The result is a corpus of Old and Middle Hungarian texts, enriched with a normalized version and morphological annotation. The rich morphology of Hungarian makes the effective retrieval of examples of many specific syntactic phenomena possible even if only morphosyntactic annotation is available.

In the sections below, we first describe the components of the corpus, followed by a description of the segmentation and normalization process and the arising issues that made these steps non-trivial. This is followed by sections about how the morphological analyzer was adapted to the task of analyzing these texts, the problems we encountered and how they were solved. We also present the automatic and the manual disambiguation system used for the morphosyntactic annotation of texts and the corpus manager with the help of which the annotated corpora can be searched and maintained. In addition, we discuss the relation of our annotated corpus to other annotated historical corpora.

2. Corpus design

2.1. Sources and resources

The overwhelming majority of extant texts from the Old Hungarian period are codices, mainly containing texts translated from Latin (in most of the cases, copies of lost translations). Another project, launched in 2009, a year prior to ours, focused on such texts[1] (Simon 2014). These are invaluable sources for investigating the history of

[1] http://omagyarkorpusz.nytud.hu/en-descr.html

Hungarian. However, it also seemed important to present a different register, the significance of which is evident in historical linguistics: informal language use. Obviously, there is no way to access spoken language, the most prototypical type of language use in this register, directly, but there are sources that could be regarded as the best possible approximation. Therefore, we decided to compile a corpus that contains minutes taken at court trials, such as witch trials, and letters sent by noblemen and serfs (the former being mostly private correspondence), in a similar vein as e.g. the Corpus of English Dialogues[2], The Corpus of Early English Correspondence[3], or the P.S. Post Scriptum Project focusing on Early Modern Portuguese and Spanish letters[4]. Although they were edited, court records are written versions of spoken testimonies of witnesses. Due to the specific nature of these trials, witnesses were warned to recall faithfully whatever they knew that would pertain to the given case, which means that they recited old conversations, threats, disputes etc. As for private letters, these are "apparently the most dialogic and interactional in the sense that the writer-addressee dyad can be located in specific individuals" (Pahta et al. 2010:7).

During the first centuries after the foundation of the Hungarian State (1000 AC), the official language was Latin. The primary language of private correspondence was also Latin. The first extant private letter in Hungarian was written only at the very end of the 15th century, and the first extant records of trials in Hungarian date back to the first half of the 16th century. Therefore, the number of texts that would represent the Old Hungarian period (prior to 1526, the year of the Turkish occupation and the consequent demolition of monastic culture) is relatively small compared to those representing Middle Hungarian (spoken between the symbolic boundaries of 1526 and 1772, the beginning of the Age of Enlightenment in Hungary), during which the amount of texts that survived up to now rose from decade to decade. Therefore, a corpus representing this register cannot be balanced with respect to time: it was clear from the outset that it has to include all the texts available from Old Hungarian, and a selection of texts available from Middle Hungarian.

2.2. Text selection and metadata

When selecting the texts to be included in the corpus, sociological features played an important role: the texts of the lawsuits span almost the whole period and are comprehensive geographically, covering several Middle Hungarian dialects. The date and location of the court case recorded as metadata in the database makes it possible to identify the dialect/language variant used in the given text. The social status of the speakers is also identifiable: all witnesses and those accused in witch trials were from lower social classes, they were either serfs or artisans. In addition, personal letters can be further subdivided according to the social status and gender of the sender, his/her relationship to the addressee, the social status and gender of the addressee, and whether the letter was written by the sender himself/herself or it was created by someone else under his/her name. These variables were recorded as metadata in the corpus facilitating sociolinguistic research concerning the Middle Hungarian epoch. Although location is part of the recorded metadata also for correspondence, this data is less relevant from a sociolinguistic point of view, because, in contrast to court records, the dialect cannot be identified based on the location where the text was dated, as most of the letters were authored by people with high mobility: noblemen and university students. Due to the nature of historical sources, it could not be among the objectives to build a corpus

[2] http://www.helsinki.fi/varieng/CoRD/corpora/CED/index.html
[3] http://www.helsinki.fi/varieng/domains/CEEC.html
[4] http://ps.clul.ul.pt

that is balanced and representative, as we faced similar problems as other teams building historical corpora: “Precisely defining the total target population, which is crucial for representativeness (...), is almost impossible for past periods with any reasonable degree of statistical validity. The texts transmitted to the present represent a random subsample of the whole population, due to largely extra-linguistic accidents. Thus, historical corpora can never even remotely capture the full variety of language” (Claridge 2008: 247., cf. Meyer 2002:37). As Hunston (2008: 156) phrased it: “All corpora are a compromise between what is desirable, that is, what the corpus designer has planned, and what is possible.” The only viable approach seemed to be to gather texts that are available (also meaning that they are not subject to copyright issues), and code as much information as possible concerning their sociolinguistic features for the user to utilize.

At present, the size of the corpus available on-line is approximately 5.8 million characters[5], more than 750,000 tokens. 52.5% consists of court records, while 47.5% is correspondence. The number of individual documents in the corpus is 1544. The part of the corpus that contains fully manually checked morphological annotation consists of more than 570,000 tokens. Moreover, the corpus is continuously being extended, as described in the Conclusion section of this paper[6].

2.3. Annotation

The corpus we present in this paper is the first historical corpus of reasonable size for Hungarian that contains full morphosyntactic annotation of all the texts included in it. The annotation procedure is described in detail in Sections 3 to 5. Following digitization, all the texts were normalized and segmented into clauses manually (see Section 3). This was followed by automatic morphological analysis and automatic disambiguation (see Sections 4 and 5). The automatically created morphological annotation (including lemmatization, the annotation of morphosyntactic features including ones often considered derivational, and segmentation of compounds) was checked and corrected manually using a web-based annotation interface (see Section 5.1) and by performing queries on the database specifically designed to spot annotation errors (Section 6).

3. Preprocessing of the texts

3.1. Digitization

Fortunately, the published sources of Middle Hungarian are ample and easy to access, and most of them are suitable for linguistic research as well. All the texts selected for our corpora were originally hand-written. However, the basis for the digitized version was always a printed edition of the texts published earlier. The printed texts were scanned and converted to a character stream using OCR. This was not a trivial task owing to the extensive use of unusual characters and diacritics. In the absence of an orthographic norm, each text applied a different set of characters; moreover, the printed publications used different fonts. Thus the only way to get ac-

[5] Although the size of corpora is in general given in tokens, we also provide this data in characters, as this would facilitate the comparison of corpora, the token number being partly determined by language type (synthetic vs. analytic languages). Punctuation marks are also often counted as independent tokens. In our case, the provided token count is the number of analyzed words in the normalized version of the corpus. This also differs from raw word count in the original texts both due to differences in orthography and because it does not include the count of foreign-language (mostly Latin) tokens in the corpus.

[6] A follow-up project of the team was started in September 2015. The focus of this project is corpus-based research on historical morphology, syntax, and, above all, variation, meaning that we will mainly exploit the corpus, but a part of the resources is allocated to further enlarging of it.

ceptable results was to retrain the OCR program[7] for most of the texts from scratch since the out-of-the-box Hungarian language and glyph models of the software did not fit the texts, especially in cases where orthography had not already been modernized in the printed edition we used. In the latter case, at least the glyph models did not need to be trained. Subsequently, all the automatically recognized documents had to be manually checked and corrected, but even so, this workflow proved to be much faster than attempting to type in the texts. We performed a single experiment at the beginning of the project comparing the time needed to type and correct a specific not-very-long document and digitizing the same document training FineReader and correcting the OCR output. The manual digitization approach took about five times as much time already for about ten pages. We did not perform exact accuracy calculations of the OCR method, but with ample OCR training, proofreaders came across relatively few errors. The amount of time spent on the complete digitization procedure of our sources including proofreading ranged from 4 to 9 weeks per volume, with proofreading not done as a full-time job.

The files that emerged through this process served as the basis of the normalized texts, that is, transcripts that the morphological analyzer can handle automatically.

3.2. Segmentation

Following digitization, the texts were segmented into clauses. Finding clause boundaries in the given text types is less problematic than finding sentence boundaries, since in the case of clauses, one can rely on certain grammatical phenomena that help segmentation (one needs to identify predicates and their arguments and adjuncts). Segmentation into sentences is more arbitrary due to complete lack or inconsistency of the use of punctuation and capitalization in the original documents. Sentence as a category serves only practical purposes in this corpus: sentences are the default units of displaying information when the corpus is queried. An automatic preliminary segmentation into clauses was performed following digitization based on regular expression-based patterns relying on punctuation, conjunctions and relative pronouns as clues. When creating the normalized versions of the texts on the basis of the original sources, it was also a task of the participants to check and correct the segmentation into clauses, and to keep the segmentation consistent between the original and the normalized versions.

On sentence boundaries both the precision and the recall of this simple segmentation tool was high (P=0.95, R=0.92 evaluated on the whole corpus). The most frequent error classes include errors due to the usage of non-standard abbreviations, lack of capitalization at the beginning of sentences, usage of full stops at the end of words that are not abbreviations, the usage of question marks in the middle of interrogative clauses (especially following the question particle *-e*), and sentence-final numbers (years). 83% of all sentence boundary errors is located in parts of text written in Latin. The precision of clause boundary detection in general proved to be high (P=0.93): splits rarely needed to be undone manually. Recall on sentence-internal clause boundaries, on the other hand, was relatively low (R=0.47). The tool split clauses only if a comma was followed by an explicit conjunction. In a significant part of the texts, punctuation was missing (lack of commas) or used inconsistently.

[7] We used FineReader, which makes full customization of glyph models possible, including the total exclusion of out-of-the-box models.

The identification and systematic marking of non-Hungarian (mostly Latin) insertions was also carried out manually during this phase: these were put into curly brackets so that the parser would skip them (for further details concerning the handling of non-Hungarian insertions, see below).

3.3. Normalization

Many authors describing historical corpora highlight the importance of normalization, i.e. creating a transcript of the texts that is uniform regarding their orthography and phonology. The most obvious solution to this problem is modernization to present-day orthography, which means that the texts can be interpreted by a morphological analyzer that was developed for the modern standard version of the given language (see e.g. McEnery and Hardie 2010; Lüdeling and Kytö 2008; Bennet et al. 2010; Hendrickx and Marquilhas 2011; Archer et al. 2015). As this is a highly time-consuming, but inevitable task, different types of software were developed to assist this process, based on a manually normalized training corpus that helps to identify spelling variants (see e.g. Schneider 2002; Rayson et al. 2007; Baron et al. 2011; Archer et al. 2014, 2015; Lehto et al. 2010; Bollmann 2013). However, this did not seem to be a realistic option in our case. On the one hand, our sources displayed an enormous amount of orthographic (and partly dialectal) variation: there were almost as many systems as scribes, if there was a system at all (similar problems were encountered by other teams developing similar types of corpora, see e.g. Hendrickx—Marquilhas 2011). On the other hand, due to the rich morphology of Hungarian, an automatic normalization effort would be confronted by a much more massive amount of ambiguity and much more massive data sparseness problems than in the case of inflecting languages. The much higher morphological variety due to the combined effect of the high number of different affixes and the fact that they tend to be stacked makes a corpus in an agglutinating language much more underrepresent possible word forms than a similar-sized corpus in an inflecting language.[8] Moreover, there were a huge number of instances when team members needed to consider carefully whether a given part of a character string is a morpheme or only an accidental spelling variant. Thirdly, this decision was also motivated by the characteristics of the team, consisting mainly of historical linguists, who were willing to do this task manually.

During the process of normalization, certain phonological dialectal variations were neutralized. The main principle of normalization seemed to be simple: exhaustive representation of the original morphological structure. No extinct morphemes were replaced by their present day counterparts. We also retained extinct allomorphs unless the variation was purely phonological. In practice, however, this turned out to be a much more complicated issue, owing to the amount of variation characteristic of the sources, some of which is due to language change in progress. Still, irresolvable ambiguity had to be handled somehow, meaning that we needed to find the method that would yield an acceptable input for the morphological analyzer while not obscuring the fact that a given form may have more than one reading. In order to provide a concise description of this process, below is a list containing the typical instances where we had to make a general decision whether to normalize or not to normalize:

[8] I.e. the chance that the next token in the corpus differs from all previous tokens is and remains much higher for any corpus size.

A) What we did normalize:

- purely orthographic variation (*mondgya mongya mondga mondgja mongja mongia mondgia* → *mondja* ‘says’*)*
- purely dialectal phonological variation (*pöcsétünkel* → *pecsétünkkel* ‘with our seal’)
- names (*Ersik Ersok Ersek* → *Erzsók* ‘Betty’)
- instances of code-switching to Latin that were integrated into the Hungarian text (*occurál* → *okkurál* ‘occur’)

B) What we did not normalize (see details below):

- forms that carry relevant morphological information from a historical point of view
- instances of code-switching to Latin that were not deemed to be integrated into the Hungarian text
- forms that are ambiguous due to ongoing change and/or orthographical ambiguity

These decisions were made in order to give the best possible input for the morphological analyzer. For instance, even though the orthography of names showed a considerable variation, it was necessary to normalize these as well (as opposed to the solution applied by Hendrickx and Marquilhas 2011, tagging them as ‘name’ as they are), because they are inflected in the majority of the cases, and for an automatic morphological analysis to be performed on them, a stem with some consistent spelling needs to be identified. Similarly, instances of code switching had to be handled sensibly. For instance, it was quite regular in the era to apply Hungarian inflection to Latin verb stems, which naturally means that they needed to be parsed. On the other hand, there were numerous Latin items that did not show overt Hungarian morphological marking, yet this can also be due to their morphosyntactic properties (nouns in the nominative case, which is unmarked in Hungarian, or adverbials), therefore the lack of overt morphological marking cannot be taken as a clue to identify different types of code-switching (alternational vs. insertional). Besides, while we did not want to leave significant constituents of basically Hungarian clauses unanalyzed, we did not want to undertake the analysis of larger chunks of foreign text either. Therefore, if a larger chunk of the text that was included into the normalized version was in Latin, it was marked as foreign, and it was left unannotated by the morphological analyzer. However, if a given item was considered an integral part of the Hungarian clause, it was inserted into the normalized version as text to be analyzed.

Concerning ambiguity, sometimes the simplest decision, i.e. whether to write a letter with or without an accent mark on it, or whether to write two items as one word or two, led to lengthy discussions and specific solutions, as these seemingly meticulous distinctions affect morphological analysis. In what follows, we will survey typical instances of ambiguity that required specific solutions already at the phase of normalization. These cases are heterogeneous with respect to the result of the chosen normalization practice: some of these determined the morphological analysis of the given item themselves, whereas in certain cases it was necessary to combine a given normalization strategy with a specific annotation procedure.

3.3.1. Masking ambiguity during normalization

The class of preverbal particles is a typical instance of having no other option but masking ambiguity during normalization. The category of preverbs was constantly growing during the given era, and the source of this grammaticalization process was the class of adverbials and postpositions. In Modern Standard Hungarian

(MSH), preverbs and verbs are written as one word, whereas neither adverbials, nor postpositions form a compound-like structure with the verb orthographically. Therefore, if one finds an ambiguous structure, this decision (i.e. whether to write it as one or more words in the normalized version) excludes the alternative morphological analysis. There was therefore no other option but to define a rule of thumb: in case of adverb–preverb ambiguity, the two items were written as one word, and the user is reminded of the possibility of an alternative morphological analysis in the user's guide (similar problems were encountered by other teams, see eg. Bennet et al. 2010). To tell the truth, distinctions of this kind would not reflect significant semantic distinctions, but such a disclaimer is appropriate.

Another case when there was no other possibility but to hide morphological ambiguity during normalization was that of the definite article. Although introduction of a definite article was an innovation of Old Hungarian, its form (which is *a/az* in MSH depending on whether the following word is consonant- or vowel-initial) was not yet consistently differentiated from that of its source, the distal deictic pronoun *az* ('that'). Therefore, the form *az házat* could have two interpretations: 'the house.acc' or 'that house.acc', the first being *a ház-at*, the second *az-t a ház-at* in MSH. In addition to the problem of two possible interpretations, the normalized version of the second reading would yield a form that consists of morphemes not present in the original version. The solution in this case was similar to the previous one, a rule of thumb saying that in these cases, one has to normalize the ambiguous form as if it were a definite article (which, in turn, already determined its morphological analysis), and the user will be reminded of the other possibility in the user's guide. This is the method we applied if ambiguity pertains to a given class of words as a whole.

3.3.2. Marking morphological ambiguity during normalization with an associated morphological annotation procedure

Another type of ambiguities required specific solutions both during normalization and morphological annotation. In these cases, the given normalization instruction was paired up with a complementary annotation procedure (for those cases in which this also required the extension of the tag set of the analyzer, see the next section). For instance, the suffixes marking the inessive (*-bAn*) and the illative (*-bA*) were used in different orthographic traditions in the Middle Hungarian era with different distribution than in written MSH. (As a matter of fact, there is great variation in this respect even in modern *spoken* Hungarian). In cases where the usage in the text differs from written MSH, we kept the form present in the text, but deviation from MSH orthography in these cases was marked by an apostrophe (e.g. *házba'n*, if the original version of the text had a locative form where the context would suggest a lative reading 'house.Ill', and *házba'*, if the original version of the text had a lative form where the context would suggest a locative reading 'house.Ine'). The morphological analyzer assigns the analysis appropriate in the given context to these forms (i.e. illative to forms normalized as *-bA'n* and inessive to forms normalized as *-bA'*). This way, these forms can be easily searched for using a query referring to both their form and their function.

An example of irresolvable ambiguity due to the vagueness of the orthography of the era is the non-consistent marking of vowel length. Due to accents missing from the texts, definite and indefinite 3rd person singular imperfect verb forms were often not distinguished, e.g. those of the frequently used word *mond* 'say': *mondá* ~ *monda*. Furthermore, in many texts in the corpus, these two forms were used with a clearly different distribution from their MSH (3rd person singular past) counterparts *mondta* ~ *mondott*. Therefore, in many cases, neither the orthography, nor the usage was consistent enough to decide unambiguously how a certain appearance of *monda* should be interpreted concerning definiteness. In these cases, we used flying accents (*monda´*) during normalization.[9] Forms marked like this are analyzed by the morphological analyzer as inherently ambiguous concerning the definiteness of the object of the verb. Incidentally, 1st person singular past forms (e.g. *mondtam* 'I said') or 2nd person plural imperfect forms (*mondátok* 'you.pl said') are similarly ambiguous concerning definiteness, and this ambiguity is not due to orthographical issues, but is simply a case of paradigmatic syncretism. Nevertheless, disambiguation of these forms is not always possible based on the present-day grammatical intuition of a human annotator, because we know that it may in specific cases differ from that of the author of the original text. Nevertheless, the latter forms can be marked as inherently ambiguous only during morphological analysis and disambiguation, while the forms ambiguous due to inconsistent marking of vowel length must be marked as such already at the time of normalization.

The vagueness of orthography may result in even more ambiguity than in the case of *monda*: the sequence *halla* could either be interpreted as indefinite 3rd person singular imperfect of *hall 'hear'* (*halla*), or definite 3rd person singular imperfect (*hallá*), or, due to indeterminacy of the marking of the palatality of the consonant, even definite 3rd person singular imperative, or definite 3rd person singular present indicative (both to be normalized as *hallja).* As this is a rather unique case involving only a handful of not very frequent word forms, and the ambiguity concerns a complicated pattern of possible feature combinations instead of just a single feature, the solution above is not applicable as it is. In such cases, the most probable normalized form has to be chosen (which could be ambiguous itself e.g. *halla´),* and that choice would constrain possible morphological analyses, but in such cases the user is reminded of the further potential ambiguity (i.e. that this could also be a case of a present tense definite word form, *hallja*) using an asterisk: e.g. **halla´.*

Another example of inherent ambiguity is a dialectal variant of possessive marking, which is very frequent in this corpus and often neutralizes singular and plural possessed forms. For example, the original word form *cselekedetitüll* could both mean 'due to his/her deed' or 'due to his/her deeds', which in many cases cannot be disambiguated based on the context even for human annotators. The solution in this case was to use a special normalization method, namely to normalize these forms in a manner that would not be used in Modern Standard Hungarian, e.g. as *cselekedetitől* in the given case. (In MSH, this form does not exist: 'due to his/her deed' is *cselekedetétől,* and 'due to his/her deeds' is *cselekedeteitől).* These special, non-standard forms were assigned special tags by the morphological analyzer (see the next section). In this specific case, the tag used (`PxS3.Pl?=i`) conveys the information that the given morphosyntactic feature (here number) is ambiguous (`Pl?`) and the fact that it takes a non-standard neutralized form (`=i`).

In the case of a specific group of words (containing mostly adverbials and adverbial pronouns), simple normalization to the modern standard form would not have strictly meant loss of morphological structure, but loss of valuable (diachronic) morphological information. For instance, the adverbial pronoun meaning 'where' had two alternative forms, *ahon* and *ahol*. Although they were probably already morphologically non-transparent in Middle Hungarian, the difference between them cannot be treated as purely dialectal difference, as they preserve

[9] Besides, to facilitate text input, we also allowed the use of asterisks, which are easier to type, and were subsequently converted to flying accents before morphological analysis.

different suffixes (i.e. different morphological structure), and it did not seem to be a very good practice to neutralize this difference by normalizing them both to *ahol*, the modern standard form. The solution was to maintain both forms in the normalized version, and to co-index their lemmas. In practice this means that if users search for the occurrences of *ahol*, i.e. the modern standard form, they will get a list containing both the occurrences of *ahol* and *ahon*. From then on, they can choose to filter out the examples of these latter forms if these data would be considered irrelevant. Hopefully, this will prove to be a user-friendly feature, as this facilitates the search for morphological fossils.

The segmented normalized version, i.e. the input of the morphological analyzer, looks like the examples in Figure **1** (the lines in boldface come from the original version of the text, below them there are their normalized versions, i.e. those that approximate Modern Standard Hungarian). Note that modernization to present-day orthography also implies differences in tokenization into individual words between the original and the normalized version. This was accomplished using special marking of white space where there is a mismatch in tokenization (marked with a backslash, see e.g. Figure **1**, last example).

en tiltottam Én tiltottam,	*I banned*
hogj meg ne egje hogy meg ne egye.	*that she should not eat it.*
nekem monta Szabo Györgjne Nekem mondta Szabó Györgyné:	*Mrs. György Szabó told me:*
halgas te kutyaba telelt „Hallgass, te kutyába' telelt,	*Shut up you who wintered in a dog,*
nem tucz te ahoz nem tudsz te ahhoz!"	*what do you know to that?*
a Menyecske meg\ even a gjökeret, A menyecske megevén a gyökeret,	*The young woman having eaten the root*
kerte tüle kérdte tőle,	*she asked her*
joé jó-e.	*if it was good.*
Menyecske felelt Menyecske felelt,	*Young woman replied:*
eleg edes: elég édes.	*It is sweet enough*
Haza\ menven a Menyecske meg\ betegedet, Hazamenvén a menyecske megbetegedett.	*Having gone home the young woman fell ill.*

Fig. 1 Segmentation into clauses of a fragment of Bosz. 41. Bihar county, Nagykereki, 1724[10]

The use of the backslash [\], as shown in the examples, is motivated by practical reasons: it marks cases when an item would be written as one word in Modern Standard Hungarian, but the original source contains these as two separate words. In order to achieve correct matching of the original and the normalized versions of the texts, every clause has to contain the same number of tokens; therefore this mismatch (and its inverse, when the origi-

[10] Bosz. = Schram, Ferenc (ed.), *Magyarországi boszorkányperek 1–3. [Witch trials in Hungary Vol. 1–3].* Akadémiai Kiadó, Budapest, 1983.

nal source contains one word where the normalized text has two) has to be marked systematically. Similarly, several symbols were introduced to mark if a text segment is incomplete ({\…}) or is non-Hungarian (e.g. _{\!lat!29. Septembris 698.}_), if it is or has a deleted (e.g. _Vörös {\uy} Vÿz_) or an inserted part (e.g. _Chris{%t}ina_), or if several readings are possible (e.g. _megláta,_ → *_meglátta_*). The normalized texts were automatically checked for eventual tokenization mismatches, which were then manually corrected.

Although the normalized versions of the texts were created individually, normalization as such involved teamwork. On the one hand, the project participants making the normalized versions were encouraged to consult with the others during regular meetings if they found problems they could not tackle with the help of the guidelines provided for the process of normalization. On the other hand, all text versions (i.e. the digitized originals and their normalized versions) were triple-checked by members of the team in order to minimize the number of mistakes. Naturally, if normalization is done manually by different people, there is a considerable risk that the normalized version of the texts will not be totally homogenous. In order to avoid this, the normalization guidelines were continuously updated when controversial issues were identified to include both the description of the problematic cases and the chosen solution. In some cases, already normalized (and analyzed) texts had to be postcorrected using automatic regular-expression-based query and update methods, when the final decision on the solution to a specific normalization/annotation issue was made.

4. Morphological analysis

The digitized and normalized texts have been analyzed with an extended version of the Humor analyzer for Hungarian. The lexicon of lemmas and the affix inventory of the program were augmented with items that have disappeared from the language but are present in historical corpora. The morphological analyzer was extended to handle both Old Hungarian and Middle Hungarian morphological constructions. The lemma database had to be supplemented with more than 5000 lemmas, the affix inventory with 50 new affixes (not counting their allomorphs), and more than 440 lines of rules (about 20%) were added or modified in the grammar of the morphological analyzer to handle morphological constructions not attested in MSH.

Certain affixes have not disappeared, but their productivity has diminished compared to the Old Hungarian era. Although words with these morphemes are still present in the language, they are generally lexicalized items, often with a changed meaning. While lexicalized forms containing such suffixes were present in the morphological analyzer for MSH, they had to be included as productive suffixes in the version of the morphology meant to annotate historical texts.

One factor that made adaptation of the morphological model difficult was that there are no reliable accounts on the changes of paradigms. Data concerning which affix allomorphs could be attached to which stem allomorphs had to be extracted from the texts themselves. Certain morphological constructions that had already disappeared by the end of the Old Hungarian era were rather rare (such as some participle forms) and often some items in these rare subparadigms have alternative interpretations. This made the formal description of these paradigms rather difficult.

As mentioned in Section 3.2, a number of affixes needed to be added that represent inherently ambiguous forms. In these cases, a question mark in the tag belonging to the affix marks that the word form is ambiguous concerning the grammatical feature denoted by the tag, e.g. *mondtam*{mond[V.Past.S1.Def?]} ‘I said (Def?)’, kezivel{kéz[N.PxS3.Pl?=i.Ins]} ‘with his hand(s)’, monda´{mond[V.Ipf.S3.Def?]} ‘he said (Def?)’.

However, the most time consuming task was the enlargement of the stem inventory. Beside the addition of a number of new lemmas, the entries of several items already listed in the lexicon of the present-day analyzer had to be modified. The causes were various: some roots now belong to another part of speech, or in some constructions they had to be analyzed differently from their present analysis.

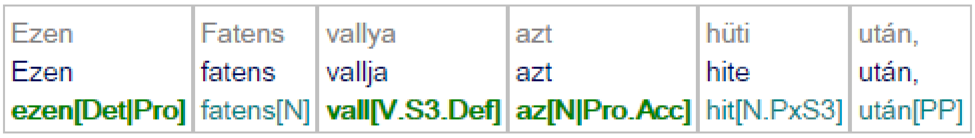
Furthermore, the number of pronouns was considerably higher in the examined period than today. The description of their extensive and rather irregular paradigms was really challenging as some forms were underrepresented in the corpora.

Some enhancements of the morphological analyzer made during the corpus annotation projects were also applicable to the morphological description of standard modern Hungarian. One such modification was a new annotation scheme applied to time adverbials that are lexicalized suffixed (or unsuffixed) forms of nouns, like *reggel* ‘morning/in the morning’ or *nappal* ‘daytime/in daytime’, quite a few of which can be modified by adjectives when used adverbially, such as *fényes nappal* ‘in broad daylight’. This latter fact sheds light on a double nature of these words that could be captured in an annotation of these forms as specially suffixed forms of nouns instead of atomic adverbs, an analysis that is compatible with X-bar theory (Jackendoff, 1977).

The morphological tagset of the original Humor analyzer is a proprietary system of Hungarian category name abbreviations. We extended this tag set to cover all the morphological constructions in Old and Middle Hungarian and internationalized it to make the tags interpretable to the international community. Nevertheless, the morphological tags used in the system do not exactly follow an international standard. There is some overlap with the tags suggested in the Leipzig Glossing Rules (LGR), but the latter covers only a fraction of the morphological features we use (not only for annotating Middle and Old Hungarian but also in the tag set of the original Humor analyzer for MSH), and there are also differences in the abbreviations the two systems use for morphosyntactic features covered by both annotation schemes. In 2016, an LGR-based complete annotation system for MSH was developed (Novák et al. 2017). We consider updating the annotations in the corpus to this scheme in the future.

5. Disambiguation

The morphological annotation had to be disambiguated. The workflow for disambiguation of morphosyntactic annotation was a semi-automatic process: an automatically pre-disambiguated version of each text was checked and corrected manually. We used a statistical HMM tagger for automatic disambiguation that was incrementally trained on the corpus itself as it grew. The tools and methods we used for automatic disambiguation as well as their performance are described in detail in section 5.2. The initial portion of the annotated training data for the tagger was created completely from scratch using the manual disambiguation interface described in Section 5.1. After the process of morphological analysis, the output is a version in which already three lines correspond to a single clause: the original text, the normalized version and the morphological annotation (the lemma followed by the morphological tag in brackets), as shown in Figure 2.



Ezen	Fatens	vallya	azt	hüti	után,
Ezen	fatens	vallja	azt	hite	után,
ezen[Det|Pro]	fatens[N]	**vall[V.S3.Def]**	**az[N|Pro.Acc]**	hit[N.PxS3]	után[PP]

‘This witness testifies that according to his faith’

Fig. 2 A sample after automatic morphological annotation, before manual disambiguation (Bosz. 273. Sopron county, Szil, 1737). The annotation of ambiguous word forms is green and bold.

The ambiguity rate of the output of the extended morphological analyzer on historical texts is higher than that for the standard Humor analyzer for present-day corpora (2.21 vs. 1.92[11] analyses/word with an identical (high) granularity of analyses). However, in some extreme cases the number of possible analyses can be much higher. The corpora contain instances of various now extinct participial and passive constructions, the addition of which to the morphological analyzer increased the number of possible analyses for some quite frequent verb forms, such as the one in Figure 3, rather significantly. This ambiguity is due to several factors:

(i) the historical analyzer is less strict (allowing now substandard or dialectal constructions, many of which coincide with regular forms),

(ii) there are several identical members of the enlarged verbal paradigms including massively ambiguous subparadigms like that of the passive and the factitive[12] and the numerous participial forms,

(iii) many inherent ambiguities described above.

[11] Measured on newswire text.

[12] This ambiguity is absent from modern standard Hungarian because the passive is not used any more.

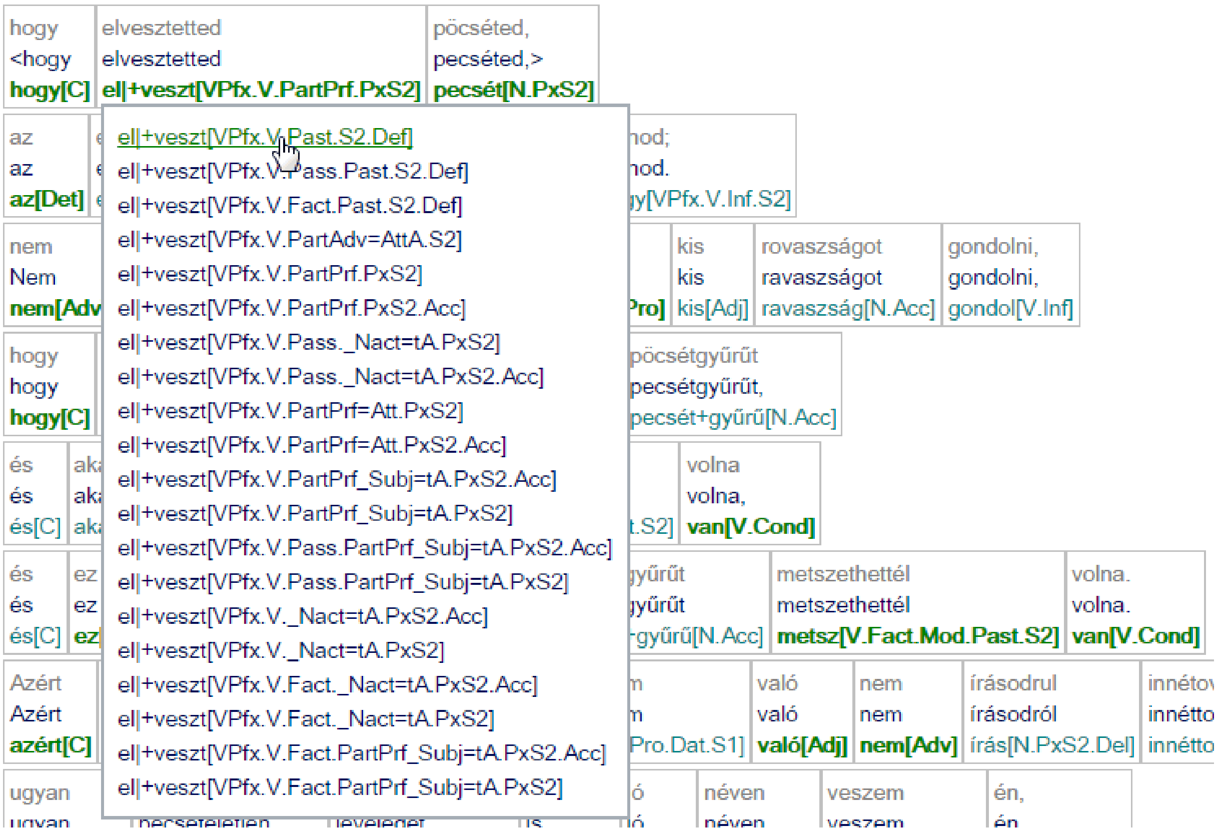

hogy
<hogy
hogy[C]
elvesztetted
elvesztetted
el|+veszt[VPfx.V.PartPrf.PxS2]
pöcséted,
pecséted,>
pecsét[N.PxS2]
el|+veszt[VPfx.V.Past.S2.Def]
el|+veszt[VPfx.V.Pass.Past.S2.Def]
el|+veszt[VPfx.V.Fact.Past.S2.Def]
el|+veszt[VPfx.V.PartAdv=AttA.S2]
el|+veszt[VPfx.V.PartPrf.PxS2]
el|+veszt[VPfx.V.PartPrf.PxS2.Acc]
el|+veszt[VPfx.V.Pass._Nact=tA.PxS2]
el|+veszt[VPfx.V.Pass._Nact=tA.PxS2.Acc]
el|+veszt[VPfx.V.PartPrf=Att.PxS2]
el|+veszt[VPfx.V.PartPrf=Att.PxS2.Acc]
el|+veszt[VPfx.V.PartPrf_Subj=tA.PxS2.Acc]
el|+veszt[VPfx.V.PartPrf_Subj=tA.PxS2]
el|+veszt[VPfx.V.Pass.PartPrf_Subj=tA.PxS2.Acc]
el|+veszt[VPfx.V.Pass.PartPrf_Subj=tA.PxS2]
el|+veszt[VPfx.V._Nact=tA.PxS2.Acc]
el|+veszt[VPfx.V._Nact=tA.PxS2]
el|+veszt[VPfx.V.Fact._Nact=tA.PxS2.Acc]
el|+veszt[VPfx.V.Fact._Nact=tA.PxS2]
el|+veszt[VPfx.V.Fact.PartPrf_Subj=tA.PxS2.Acc]
el|+veszt[VPfx.V.Fact.PartPrf_Subj=tA.PxS2]

Fig. 3 Possible analyses of a verb form (*elvesztetted* 'you lost (definite object)' before manual disambiguation) (Nád. p. 18 1557-06-24[13])

5.1. Manual disambiguation

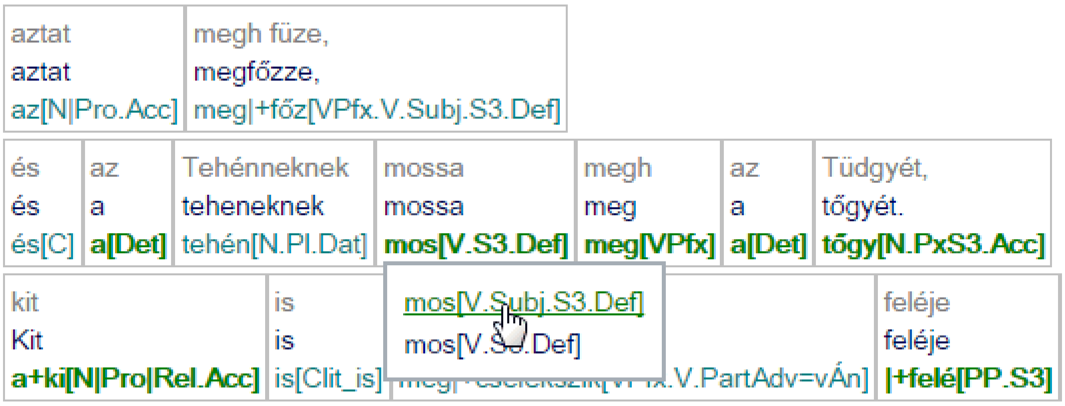
To support the process of manual validation and the initial manual disambiguation of the training corpus a web-based interface was created using JavaScript and Ajax[14] where disambiguation and normalization errors can be corrected effectively. The system presents the document to the user in an interlinear annotation format that is easy and natural to read. An alternative analysis can be chosen from a pop-up menu containing a list of analyses applicable to the word that appears when the mouse cursor is placed over the word in question. Note that the list only contains grammatically relevant analyses for the word returned by the morphological analyzer running on the web server. This is important, since, due to the agglutinating nature of Hungarian, there are thousands of possible tags and lemmatization is non-trivial (see Figure 4).

The original and the normalized word forms as well as the analyses can also be edited by clicking them, and an immediate reanalysis by the morphological analyzer running on the web server can be initiated by double clicking the word. We use Ajax technology to update only the part of the page belonging to the given token, so the update is immediate. Afterwards, a new analysis can be selected from the updated pop-up menu.

[13] Nád. = Károlyi, Árpád – Szalay, József (eds.), *Nádasdy Tamás nádor családi levelezése [Family correspondence of Tamás Nádasdy, palatine of Hungary]*. Akadémiai Kiadó, Budapest, 1882.

[14] Ajax (Asynchronous JavaScript and XML) is a client-side browser script that communicates to a server/database without the need for a complete web page refresh.

As there is an inherent difference between the original and normalized tokenization, and because, even after thorough proofreading of the normalized version, there may remain tokenization errors in the texts, it is important that tokens and clauses can also be split and joined using the disambiguation interface.



aztat	megh füze,
aztat	megfőzze,
az[N|Pro.Acc]	meg|+főz[VPfx.V.Subj.S3.Def]

és	az	Tehénneknek	mossa	megh	az	Tüdgyét,
és	a	teheneknek	mossa	meg	a	tőgyét.
és[C]	**a[Det]**	tehén[N.Pl.Dat]	**mos[V.S3.Def]**	**meg[VPfx]**	**a[Det]**	**tőgy[N.PxS3.Acc]**

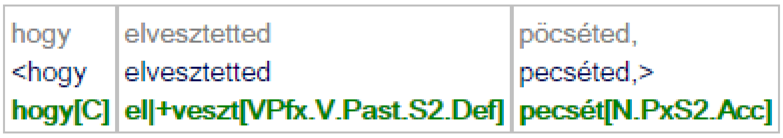
kit	is		feléje
Kit	is		feléje
a+ki[N|Pro|Rel.Acc]	is[Clit_is]	V.PartAdv=vÁn]	**|+felé[PP.S3]**

mos[V.Subj.S3.Def]
mos[V.S3.Def]

Fig. 4 The web-based disambiguation interface (‘… she should cook that/ and wash the udders of the cows./ Having done that toward her…’ – Bosz. 192. Pest county, Veresegyház, 1744.)

In cases when the morphological analyzer assigns several possible analyses to one form, the program selects the most likely analysis automatically (see Section 5.2), but these cases are highlighted (displayed in green instead of blue), and the annotators need to check whether the choice of the program is correct. In this phase of the project, it is a time-consuming, but inevitable task to check and disambiguate these cases manually, i.e. to select the analysis that fits the context. Figure 3 shows an instance of a highly ambiguous verb form in context that was not correctly disambiguated by the system, while Figure 5 shows the same clause with the given verb form correctly disambiguated.

The abundance of ambiguous forms highlighted (boldface and green in Figures 4 and 3) shows that many forms need to be checked manually. However, as the automatic disambiguation tool is incrementally retrained (manual disambiguation serving as feedback), its “guesses” are getting more and more reliable. Moreover, the more familiar annotators get with the morphological codes used in the annotation, the less time and effort they need to spend on spotting erroneous annotation.



hogy	elvesztetted	pöcséted,
<hogy	elvesztetted	pecséted,>
hogy[C]	**el|+veszt[VPfx.V.Past.S2.Def]**	**pecsét[N.PxS2.Acc]**

Fig. 5 A disambiguated string (‘that you lost your seal’ – Nád. p. 18 1557-06-24)

Another fairly labor-intensive task was the checking, correction and extension of the database of the morphological analyzer itself, which required constant cooperation between members of the team. Of the many problems that needed to be solved, a characteristic type was when the tags used by the morphological analyzer had to be fine-tuned. For instance, we needed to add a new type of analysis for word forms heading a type of postpositional phrase construction present in the texts but no longer existing in MSH. This, naturally, leads to a dilemma: increasing the number of tags yields more reliable and precise morphological analysis, but potentially raises difficulties if a user wants to compare data taken from different corpora. No better solution seemed to emerge to this problem than to inform the user as thoroughly as possible on the function of the tags in the user’s guide.

The automatic annotation system was designed to facilitate the modification of details of the annotation scheme in the course of work. One such modification was e.g. the change to the annotation of time adverbs mentioned above. The modified annotation can be applied to texts analyzed and disambiguated prior to the change of the annotation scheme relatively easily. This is achieved by the fact that, in the course of reanalysis, the program chooses the analysis most similar to the previously selected one (based on a letter trigram similarity measure). Nevertheless, the system highlights all tokens the reanalysis of which resulted in a change of annotation, so that these spots can be easily checked by the annotators. For changes in the annotation scheme where the simple similarity-based heuristic could not be expected to yield an appropriate result (e.g. when we decided to use a more detailed analysis of derived verb forms than before), a more sophisticated method was used to update the annotations: old analyses were replaced using regular expressions generated automatically. These were created based on a manually checked output of the morphological generator. The simple internal annotation format we use (see Section 5.3) made the definition of these regular-expression-based replacements relatively easy.

5.2. Automatic disambiguation

The first few documents were disambiguated completely manually using the web-based tool. Later we started to train and use a tagger for pre-disambiguation retraining the tagger incrementally on an increasing amount of disambiguated and checked text. First we used the HMM-based trigram tagger HunPos (Halácsy et al. 2007). HunPos is not capable of lemmatization, but we used a straightforward method to get a full analysis: we applied reanalysis to the text annotated only with the tags assigned by the tagger using the automatic trigram-similarity-based ranking of the analyses. This approach yielded rather good results, but one problem with it was that the similarity-based ranking inevitably prefers shorter lemmas. This was not appropriate for handling a frequent type of lemma ambiguity for Hungarian verbs with one of the lemma candidates ending in an *–ik* suffix and the other lacking that suffix (such as *dolgozik* ‘work’ vs. *(fel)dolgoz* ‘process’). Always selecting the shorter *–ik*-less variant is not a good choice in the case of many frequent words in this ambiguity class.

Later, HunPos was replaced with another HMM-based trigram tagger, PurePos (Orosz – Novák 2012), that has some attractive extra features. It can process morphologically analyzed ambiguous input and/or use an integrated analyzer constraining possible analyses to those proposed by the analyzer or read from the input instead of relying on a suffix guesser for words not seen in the training corpus. This boosts the precision of the tagger dramatically in the case of languages like Hungarian and small training corpora. The fact that PurePos can be fed analyzed input makes it easy to combine with constraint-based tools that can further improve the accuracy of the tagging by handling long distance agreement phenomena not covered by the trigram model, or by simply removing impossible tag sequences from the search space of the tool.

PurePos can also perform lemmatization, even for words unknown to the morphological analyzer (and not annotated on the input) learning a suffix-based lemmatization model from the training corpus along with a similar suffix-based tag guessing model, thus it assigns a full morphological analysis to every token. It is also capable of generating an n-best list of annotations for the input sentence when using beam search instead of the default Viterbi decoding algorithm.

Automatic disambiguation performance

We performed an evaluation of the accuracy of PurePos on an 84000-word manually checked portion of the corpus using five-fold cross-validation with a training corpus of about 67000 words and a test corpus of around 17000 words in each round.[15] The ratio of words unknown to the morphological analyzer in this corpus was low: 0.32%.

The average accuracy of tagging, lemmatization and full morphological annotation for two versions of the tagger are shown in Table 1. In addition to token accuracy, we also present clause accuracy values in the table. Note that, in contrast to the usual way of evaluating taggers, these values were calculated excluding the always unambiguous punctuation tokens. The baseline HHM tagger in Table 1 uses no symbolic morphological information at all. The implementation of this model used suffix guessing for lemmatization in all cases (even for words seen in the training corpus) and selected the most frequent lemma, which is obviously not an ideal solution.

The disambiguation tool using morphologically analyzed input performed significantly better. Its clause-level accuracy was 81.50%, meaning that only every fifth clause contained a tagging error. We use a rich tag set in the corpus that differentiates constructions which are not generally differentiated at the tag level in Hungarian corpora, e.g. deictic pronouns (*ebben* ’in this’) vs. deictic pre-determiners (*ebben a házban* ’in this house’). Many instances of these words can only be disambiguated using long-distance dependencies, i.e. information often not available to the trigram tagger. Combining the tagger with a constraint-based tool (see e.g. Hulden and Francom 2012) could presumably further improve accuracy.

We listed a theoretical upper limit of the performance of the current trigram tagger implementation in the right-most column using 5-best output and assuming an ideal oracle that can select the best annotation.

		baseline HMM	with morphology	5-best+oracle
token	tag	90.17%	96.44%	98.97%
	lemma	91.52%	98.19%	99.11%
	full	87.29%	95.90%	98.53%
clause	tag	62.48%	83.81%	93.99%
	full	54.68%	81.50%	91.47%

Table 1 Disambiguation performance (accuracy) of the tagger with and without using the MA

5.3. The internal representation of the corpus

We use a proprietary format to store the corpus that was developed in 2005 when we created annotated corpora for Uralic minority languages. It is a very compact line-oriented annotation format somewhat similar in spirit to the stream format used in the Apertium machine translation system. Whitespace is retained from the original

[15] Five-fold cross-validation is an evaluation technique, where the corpus is divided into five roughly equal-sized parts. Four parts are used as a training corpus, while the fifth part is used for testing in each of the five rounds of evaluation. Results of the five evaluations are averaged.

source, except that the text is broken into separate lines at clause boundaries. Final punctuation distinguishes sentence boundaries from clause boundaries. Gapping is marked by angle brackets: an opening bracket (<) is inserted at the beginning of the clause inserted into another one, while a closing bracket (>) marks the end of the gap. Paragraphs are separated by empty lines.

Morphosyntactic annotation of each word is joined to the word in a structure enclosed in a pair of delimiters ({{ and }} by default), with alternative possible annotations separated by a separator (|| by default). The first analysis is the actual one. An asterisk marks that the analyses were actually disambiguated. The morphological analyzer can output its analyses in this format. The PurePos tagger and lemmatizer can read the output of the morphological analyzer and use the morphological analyses in its input to constrain the possible tags and lemmas for the given word form. As we have seen, this improves annotation accuracy considerably compared to that of the purely statistical suffix-based guessing models implemented in PurePos (see Table 1). In the final disambiguated version of the corpus, alternative analyses are removed. The original form of each word is separated form the normalized transcript by the separator %=.

The internal representation of the annotated fragment in Figure 5 looks like the following:

hogy%=<hogy{{*hogy[C]}} elvesztetted%=elvesztetted{{*el|+veszt[VPfx.V.Past.S2.Def]}}
pöcséted,%=pecséted,>{{*pecsét[N.PxS2.Acc]}}

Fig. 6 Internal representation of the disambiguated string in Figure 5 (‘that you lost your seal’)

Metadata and tags enclosing special structures (such as addressing or notes on the envelope in the case of letters, etc.) are stored on separate lines starting with a hash mark (#). These lines are ignored by the morphological analyzer and the tagger. We use TEI-compatible tag names to encode metadata concerning author, date, etc.

The corpus is stored in an Emdros database for querying (see Section 6). The Emdros MQL corpus generation script is generated directly from the internal corpus representation described above. The Emdros database can also be exported and converted back to the internal stream format.

A great advantage of the stream format is that it is much easier to read and to process than XML. This format makes it easy to define regular-expression-based substitution expressions that can be used to automatically fix annotation errors due either to manual or automatic operations, to standardize cases where different annotators followed different practices etc.[16] We also used this method to update annotations in the corpus whenever we decided to change certain aspects of the annotation. It is also easy to formulate expressions that refer to properties of not only the affected word but also its context. For visualization in the browser, we convert this format into HTML code that is displayed in the browser as interlinear annotation with JavaScript displaying the disambiguation pop-up boxes, as shown in Figures 3 to 5.

[16] The following is an example of a regular-expression-based substitution expression that we used to correct word forms and their analyses in which some form of the word *gyermek* ‘child’ was overnormalized to a corresponding form of *gyerek* ‘child’:

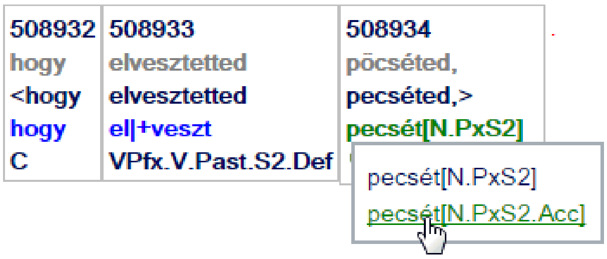
```
#gyerek > gyermek
/((?:^|\s)\S+erm\S+%=[Gg]yer)(ek\S*\{\{\*gyer)(?=ek)/$1m$2m/
```

The internal stream annotation format can also be transformed into XML easily. However, we have not yet felt the need to do so. CoNLL-U is a more likely target representation format that we will use if we decide to add syntactic annotation.

6. The query interface

The web-based corpus query tool does not only make it possible to search for different grammatical constructions in the texts, but it is also an effective annotation correction tool. Errors discovered in the annotation or the text appearing in the “results” box can immediately be corrected selecting the right annotation from the output of the morphological analyzer running on the server embedded using Ajax or editing any property of the token, and the corrected text and annotation is recorded in the database. Naturally, this latter functionality of the corpus manager is only available to team member users having the necessary privileges.

A fast and effective way of correcting errors in the annotation is using the query interface to search for presumably incorrect structures and to correct the truly problematic ones at once. Corrections are recorded in the corpus immediately. The corrected corpus can be exported after this procedure and the tagger can be retrained on it. Figure 7 shows an example of using the web interface for correction of morphological annotation errors in the returned query results. The interface makes not only the correction of individual word forms or their annotation possible but also that of errors in clause segmentation.

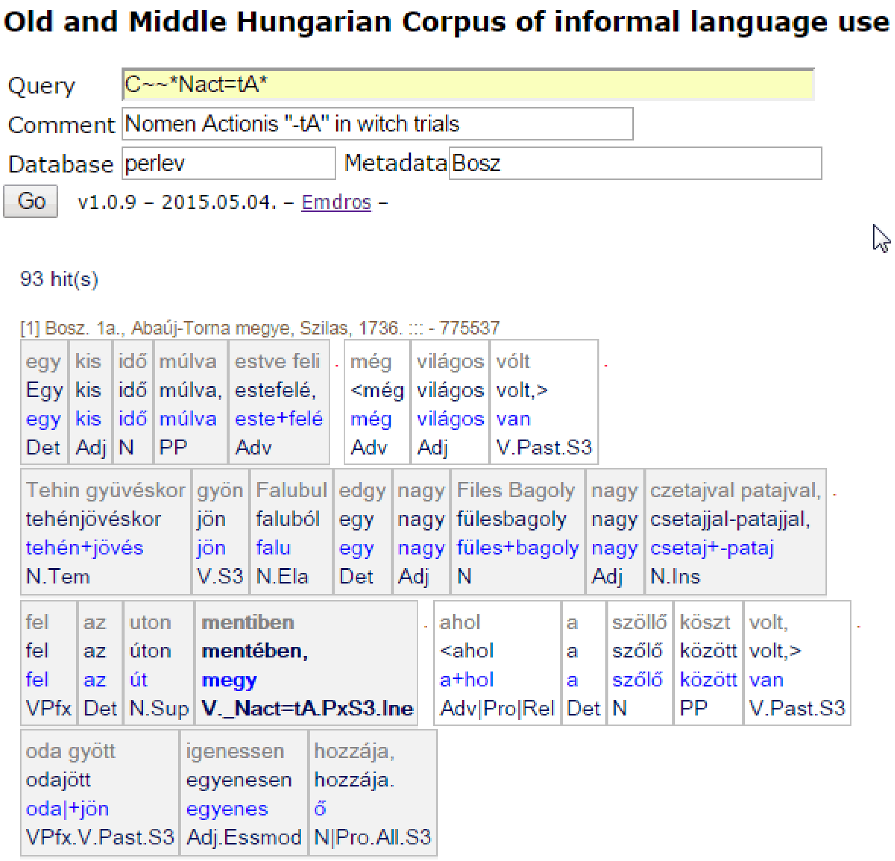


508932	508933	508934
hogy	elvesztetted	pōcséted,
<hogy	elvesztetted	pecséted,>
hogy	el|+veszt	pecsét[N.PxS2]
C	VPfx.V.Past.S2.Def	

pecsét[N.PxS2]
pecsét[N.PxS2.Acc]

Fig. 7 Manual correction of an annotation error spotted in the query results

The database used by the corpus manager is based on the Emdros corpus management and query tool (Petersen 2004). In addition to queries formulated using MQL, the standard query language of Emdros, either typed in at the query box (see Figure 8) or assembled using controls of the query interface, advanced users can use a custom-made corpus-specific query language (MEQL), which makes a much more compact formulation of queries possible than MQL does. It is e.g. extremely simple to locate a specific locus in the corpus: one simply needs to type in the sequence of words one is looking for. Queries formulated in MEQL are automatically converted to MQL queries by the query processor, which are in turn processed by the Emdros backend. Cleverly formulated MEQL queries can be effectively used to search for instances of many types of syntactic structures even though only a morphosyntactic annotation is available in the corpus.



Old and Middle Hungarian Corpus of informal language use

Query C~~*Nact=tA*
Comment Nomen Actionis "-tA" in witch trials
Database perlev Metadata Bosz
Go v1.0.9 – 2015.05.04. – Emdros –

93 hit(s)

[1] Bosz. 1a., Abaúj-Torna megye, Szilas, 1736. ::: - 775537

egy	kis	idő	múlva	estve feli	még	világos	vólt	
Egy	kis	idő	múlva,	estefelé,	<még	világos	volt,>	
egy	kis	idő	múlva	este+felé	még	világos	van	
Det	Adj	N	PP	Adv	Adv	Adj	V.Past.S3	

Tehin gyüvéskor	gyön	Falubul	edgy	nagy	Files Bagoly	nagy	czetajval patajval,
tehénjövéskor	jön	faluból	egy	nagy	fülesbagoly	nagy	csetajjal-patajjal,
tehén+jövés	jön	falu	egy	nagy	füles+bagoly	nagy	csetaj+-pataj
N.Tem	V.S3	N.Ela	Det	Adj	N	Adj	N.Ins

fel	az	uton	**mentiben**	ahol	a	szöllő	köszt	volt,
fel	az	úton	**mentében,**	<ahol	a	szőlő	között	volt,>
fel	az	út	**megy**	a+hol	a	szőlő	között	van
VPfx	Det	N.Sup	**V._Nact=tA.PxS3.Ine**	Adv|Pro|Rel	Det	N	PP	V.Past.S3

oda gyött	igenessen	hozzája,
odajött	egyenesen	hozzája.
oda|+jön	egyenes	ő
VPfx.V.Past.S3	Adj.Essmod	N|Pro.All.S3

Fig. 8 The query interface

The search engine makes it possible to constrain the scope of a search to sentences, clauses, or texts containing grammatical constructions and/or tagged with metadata matching the criteria specified in the query. Units longer than a sentence can also be searched for. The context displayed by default for each hit is the enclosing sentence with focus tokens highlighted. The whole document containing the given result can be displayed in another browser window by a single click on the reference containing the document metadata. In this document view, all focus tokens are highlighted just like in the original results. Clauses may be non-continuous. This is often the case for embedded subordinate clauses. But the corpus also contains many injected parenthetical coordinate clauses and examples where the topic of a subordinate clause precedes its main clause with the net effect of the subordinate clause being interrupted by the main clause. The query example in Figure 8 shows a sentence containing several clauses with gaps: the clauses enclosed in angle brackets are wedged within the topic or between the topic and comment part of the clauses which they interrupt. Emdros is capable of representing these interrupted clauses as single non-contiguous objects with the interrupting clause not being part of the interrupted one.

7. Relation of the corpus to other annotated historical corpora

7.1. Historical corpora of Old and Middle Hungarian

Although the corpus of Old Hungarian texts mentioned in Section 2.1 (Simon 2014) is bigger in size, it is only partially normalized (11 of the 47 codices included and some shorter texts), an even smaller part of the corpus was morphologically annotated (5 of the codices: 132378 word tokens: the currently accessible fully analyzed portion of the corpus presented in this paper is 5.7 times bigger), and only a fraction of the annotations was manually checked (3 codices: 63507 word tokens: the manually checked portion of our corpus is 8.3 times bigger)[17]. Note that the lack of normalization and annotation renders the majority of that corpus practically unsearchable.

Moreover, the Old Hungarian corpus project relied significantly on results obtained from efforts made within the framework of the project described in this paper. The morphological annotation (where available) in both corpora was created by the first author of this paper using the annotation tools described here. Morphological analysis was performed using the computational morphology described in Section 4 below, originally developed for MSH and adapted to Old and Middle Hungarian. Three of the fully analyzed five codices in the Old Hungarian corpus are from the Computational Database for Historical Linguistics (CDHL: Jakab and Kiss 1994, 2001; Jakab 2002). CDHL is a dictionary-like resource containing the original word forms of a few Old Hungarian codices, listing of the locations of their occurrences in the codex identifying page and line number, the stems in modern transcription and morphosyntactic annotation of the words. Reconstruction of the texts from this resource required a painstaking effort involving manual reconstruction of the order of words in each line of the codices. The normalized version of the texts was generated using the morphological generator based on the Old and Middle Hungarian computational morphology and then reanalyzed using the morphological analyzer to make the analyses compatible with that in the rest of the corpus.[18]

Morphological disambiguation of the other annotated codices in the corpus was performed by the HMM tagger described in Section 5.2, using statistical models created from manually checked annotations of texts in the Middle Hungarian corpus described in this paper. The automatically disambiguated annotations of these two codices were not checked and corrected manually. Since the morphological annotations in both the Old Hungarian corpus and the corpus described in this paper were created using the same tools and annotation scheme, the tags used in the two corpora are almost identical with a few exceptions. However, the guidelines for normalization and annotation were to some extent different. Later, the morphological annotations in the Old Hungarian corpus were also transformed to the format defined in the Universal Morphology annotation scheme keeping the original annotation format as well (Simon and Vincze 2016), and the five annotated codices were made available in the CoNLL-U format. However, contrary to what might be suggested by this format[19], the Old Hungarian corpus contains no syntactic analysis. Neither does our corpus. We discuss the reasons for the lack of syntactic annotation in the next session.

[17] Manual validation of the annotations was not performed within the framework of the Old Hungarian corpus project, but the disambiguated morphological annotations were taken from the Computational Database for Historical Linguistics (CDHL), see below.

[18] The original morphological annotations is CDHL are encoded in a hard-to-read numerical format, which occasionally was incorrect and often incomplete lacking some rather relevant distinctions (e.g. infinitives and all types of participles were collapsed into a single category in the original CDHL annotation.). Due to this, the original form often needed to be taken into account in addition to the morphological annotation when generating the normalized version of the corpus, and morphological analysis subsequently automatically added the missing morphological features to the annotation.

[19] The CONLL-U format is in general used to store treebanks containing dependency annotation.

7.2. Relation of the corpus to historical corpora in other languages

A number of historical corpora, such as the Penn Corpora of Historical English[20], the Tycho Brahe Parsed Corpus of Historical Portuguese[21], the Welsh Prose[22] corpus, the University of Ottawa parsed corpus of historical French[23], Icelandic Parsed Historical Corpus (IcePaHC)[24], the The Parsed Old and Middle Irish Corpus (POMIC)[25], the Parsed Corpus of Early New High German[26] and the Penn Parsed Corpora of Historical Greek (PPCHiG)[27] include an annotation of syntactic structure in addition to morphosyntactic annotation. Most of these corpora use an adapted version of the Penn Treebank annotation scheme, a constituency-based syntactic annotation scheme based on government and binding (GB) theory.

We, in contrast, decided to restrict the annotation in our corpus to morphosyntactic structure. The reasons behind this were mostly practical. Firstly, there were already tools available for morphological analysis, as there was a morphological analyzer developed for Modern Standard Hungarian (MSH), although it had to be adapted for the task. Secondly, owing to the non-configurational nature of Hungarian syntax (see e.g. É. Kiss 1987, Alberti 2006: 19–20, 44–45, 70, passim), syntactic annotation of Hungarian is a controversial issue. In fact, the only version of Szeged Treebank, the only available treebank for Modern Standard Hungarian that contains "full" syntactic analysis,[28] is a dependency treebank unlike the ones mentioned above (Vincze et al. 2010). Szeged Treebank also has a constituency-based version (Csendes et al. 2005). However, clause-level structures are completely flat in that version, and the representation lacks any traces or other null elements (there are pro-drop and zero copula in Hungarian in addition to unusually frequent elliptic structures) that a GB-style analysis would predict. Any indication of complements of non-verbal heads (including participles) is also missing from that treebank version. Neither is there any representation of the structures which are in fact configurational in Hungarian, such as topic-comment structure, or the intricate internal structure of the preverbal section of the comment part of clauses including the focus position and standard landing sites of specific types of quantified phrases.

It would have been beyond the scope of the project to elaborate the syntactic annotation guidelines and undertake the tremendous task of manually annotating the corpus with syntactic structure or to fund the development of a new syntactic parser. It would be desirable to have a syntactically annotated corpus of MSH with consensual annotation principles first, before those principles could be extended and applied to the obviously more difficult-to-interpret historical texts. In addition, a dependency-based representation seems to be a much more likely candidate for a consensual syntactic annotation system for Hungarian than a GB-style constituency representation with lots of traces like the syntactic annotation used in the historical corpora mentioned above. Moreover, making the dependency relations explicit would be inevitable even in a resource of the latter type.

[20] https://www.ling.upenn.edu/hist-corpora/

[21] http://www.tycho.iel.unicamp.br/corpus/en/index.html

[22] http://www.rhyddiaithganoloesol.caerdydd.ac.uk/en/

[23] http://www.voies.uottawa.ca/corpus_pg_en.html

[24] http://linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC)

[25] http://www.dias.ie/index.php?option=com_content&view=article&id=6586&Itemid=224&lang=en

[26] https://enhgcorpus.wikispaces.com/

[27] http://www.ling.upenn.edu/~janabeck/greek-corpora.html

[28] The set of distinct dependency relations is very much streamlined in that corpus version as well. The annotation of the rather frequent elliptic structures is also very controversial.

There is a platform that can be used for cross-linguistically consistent treebank annotation for many languages: Universal Dependencies (UD, Nivre et al. 2016). This annotation scheme was defined well after our project started, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. We consider (some variant of) this formalism as the most likely future platform for syntactically annotated historical corpora for Hungarian.

Finally, the third factor that played a significant role in our decision was that, owing to the rich morphology of Hungarian, morphological annotation alone makes possible the study of many types of syntactic phenomena. Cleverly formulated queries based on morphosyntactic features can return data that mainly consists of valid examples of the syntactic structures sought.

8. Conclusions

Theoretically speaking, the building of the corpus can be considered to be concluded once the morphological labels offered by the parser are disambiguated by the participants, and the search form (adjusted to this corpus) is up and running. In practice, however, there is always room for improvement. Quite naturally, the corpus can be extended all the time. Besides, problematic cases emerge constantly, and in a fair number of instances the guidelines and the linguistic database used by the annotation tools that help the participants need to be extended or even modified in light of these, and the texts that were normalized and annotated according to an earlier version need to be updated as well. A further objective intended to accomplish soon is the development of the corpus query tool in order to make the metadata more easily accessible. If data can be filtered according to the standard sociolinguistic factors, it will be possible to conduct variable rule analyses, i.e. assess which factor groups condition the choice from among the variants of a linguistic variable.

Actually, flexibility and documentation proved to be keywords all through the project. It was necessary to find a middle course between philological, descriptive, and diachronic adequacy on the one hand and the tools and data representation available for building a morphologically annotated corpus on the other. The constant cooperation of the historical linguists (normalization and manual disambiguation) and the computational linguist (morphological analysis, query interface) required flexibility from both sides, sometimes resulting in decisions that went contrary to previous decisions. This meant that the whole corpus had to be updated according to the new principles. Moreover, thorough documentation of the principles of normalization and annotation was necessary both for coordinating teamwork and for helping the prospective users interpreting the data. This also seems to be a prerequisite for the possible standardization of annotated historical corpora in general.

All in all, one has to reach the (fairly obvious) conclusion that the more one strives to help the user, the more time it takes to build a database. Many of the complications that arose during this process seem to originate from the nature of the sources selected, i.e. informal texts that are assumed to reflect the Middle Hungarian vernacular as closely as possible. However, the participants of the project agree that the efforts pay off: as a pioneering project both in reflecting the Hungarian vernacular of a given historical period and in developing electronic means specially adjusted for this historical corpus, the first fully normalized and annotated historical corpus of Hungarian will be of great value for both historical linguists and specialists or students of related fields.

References

Alberti, Gábor (2006), *Generatív grammatikai gyakorlókönyv III. A háttérelmélet.* [Exercises for Generative Grammar. III. Theoretical background.] PTE – Bölcsész konzorcium – HEFOP Iroda, Pécs.

Baron, Alistair; Rayson, Paul & Archer, Dawn (2011), *Quantifying Early Modern English spelling variation: change over time and genre.* In Conference on New Methods in Historical Corpora, University of Manchester. Presentation: http://eprints.lancs.ac.uk/60258/1/Presentation.pdf

Archer, Dawn et al. (2014), *Normalising the corpus of English dialogues (1560-1760) using VARD2: decisions and justifications.* In 35th ICAME conference, 2014 04 30-2014 05 04, Nottingham. Abstract: http://eprints.lancs.ac.uk/72803/

Archer, Dawn et al. (2015), *Guidelines for normalising Early Modern English corpora: Decisions and justifications.* ICAME Journal, Volume 39, DOI: 10.1515/icame-2015-0001

Bennet, Paul; Durell, Martin; Scheible, Silke & Whitt, Richard J. (2010), *Annotating a historical corpus of German: A case study.* Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards", Valletta, Malta, 18 May 2010. 64-68. http://www.ims.uni-stuttgart.de/institut/mitarbeiter/scheible/publications/lrec2010.pdf

Bollmann, Marcel (2013), *Spelling normalization of historical German with sparse training data.* In Proceedings of the Corpus Analysis with Noise in the Signal workshop (CANS 2013) http://ucrel.lancs.ac.uk/cans2013/abstracts/Bollmann.pdf

Claridge, Claudia (2008), *Historical corpora.* In Lüdeling, Anke & Kytö, Merja (eds): Corpus Linguistics. An International Handbook, Volume 1. Walter DE GRUYTER, Berlin – New York. 242–259.

Csendes, Dóra; Csirik, János; Gyimóthy, Tibor & Kocsor, András (2005), *The Szeged Treebank.* In Text, Speech and Dialogue, 8th International Conference, TSD 2005, 123–31. Springer.

É. Kiss, Katalin (1987), *Configurationality in Hungarian.* Reidel, Dordrecht & Akadémiai Kiadó, Budapest.

Halácsy, Péter; Kornai, András & Oravecz, Csaba (2007). *HunPos: An Open Source Trigram Tagger.* In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, 209–12. ACL ’07. Stroudsburg, PA, USA: Association for Computational Linguistics.

Hendrickx, Iris; Marquilhas, Rita (2011), *From old texts to modern spelling: an experiment in automatic normalisation.* JLCL 26 (2), 65–76.

Hulden, Mans & Jerid Francom (2012), *Boosting Statistical Tagger Accuracy with Simple Rule-Based Grammars.* In Nicoletta Calzolari; Khalid Choukri; Thierry Declerck; Mehmet Uğur Doğan; Bente Maegaard; Joseph Mariani; Jan Odijk & Stelios Piperidis (eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey: European Language Resources Association (ELRA).

Hunston, Susan (2008), *Collection strategies and design decisions.* In Lüdeling, Anke & Kytö, Merja (eds): Corpus Linguistics. An International Handbook, Volume 1. Walter de Gruyter, Berlin – New York. 154-168.

Jackendoff, Ray (1977), *X-bar-Syntax: A Study of Phrase Structure.* Linguistic Inquiry Monograph 2. MIT Press, Cambridge, MA.

Jakab, László (2002), *A Jókai-kódex mint nyelvi emlék: szótárszerű feldolgozásban.* Debrecen: Debreceni Egyetem.

Jakab, László, & Kiss, Antal. (1994), *A Guary-kódex ábécérendes adattára. Számítógépes nyelvtörténeti adattár.* Debrecen: Debreceni Egyetem.

Jakab, László, & Kiss, Antal. (2001), *A Festetics-kódex ábécérendes adattára. Számítógépes nyelvtörténeti adattár.* Debrecen: Debreceni Egyetem.

Lehto, Anu; Baron, Alistair; Ratia, Maura & Rayson, Paul (2010), *Improving the precision of corpus methods: The standardized version of Early Modern English Medical Texts.* In Taavitsainen, Irma & Pahta, Päivi (eds.), Early Modern English Medical Texts. Benjamins, Amsterdam. 279–290.

Lüdeling, Anke & Kytö, Merja (eds.) (2008): *Corpus Linguistics. An International Handbook.* Berlin, New York: Walter de Gruyter.

McEnery, Tony & Hardie, Andrew (last updated 2010): *Investigating the Journalism of the Seventeenth Century.* http://www.lancaster.ac.uk/fass/projects/newsbooks/default.htm

Meyer, Charles F. (2002), *English Corpus Linguistics. An Introduction.* Cambridge University Press.

Nivre, Joakim; de Marneffe, Marie-Catherine; Ginter, Filip; Goldberg, Yoav; Hajič, Jan; Manning, Christopher D.; McDonald, Ryan; Petrov, Slav; Pyysalo, Sampo; Silveira, Natalia; Tsarfaty, Reut & Zeman, Daniel. 2016. *Universal Dependencies v1: A Multilingual Treebank Collection.* In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 1659–66. European Language Resources Association (ELRA)

Novák, Attila (2003), *Milyen a Jó Humor? [What Is Good Humor Like?].* In I. Magyar Számítógépes Nyelvészeti Konferencia, 138–44. Szeged: SZTE.

Novák, Attila; Rebrus, Péter & Ludányi, Zsófia (2017), *Az emMorph morfológiai elemző annotációs formalizmusa [The annotation formalism of the emMorph morphological analyzer].* In XIII. Magyar Számítógépes Nyelvészeti Konferencia, 70–78. Szeged: SZTE.

Orosz, György & 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, 539–45. Hissar, Bulgaria.

Pahta, Päivi; Palander-Collin, Minna; Nevala, Minna & Nurmi, Arja (2010), *Language practices in the construction of social roles in Late Modern English.* In Pahta, Päivi – Nevala, Minna –Nurmi, Arja – Palander-Collin, Minna (eds.) Social Roles and Language Practices in Late Modern English, (Pragmatics and Beyond NS 195.) Amsterdam: Benjamins.

Petersen, Ulrik (2004), *Emdros — a Text Database Engine for Analyzed or Annotated Text.* In Proceedings of COLING 2004., 1190–93.

Prószéky, Gábor & Novák, Attila(2005), *Computational Morphologies for Small Uralic Languages.* In: Inquiries into Words, Constraints and Contexts, 150–157..

Rayson, Paul; Archer, Dawn; Baron, Alistair; Culpeper,Jonathan & Smith, Nicholas (2007), *Tagging the Bard: Evaluating the Accuracy of a Modern POS Tagger on Early Modern English Corpora.* In Proceedings of the Corpus Linguistics Conference: CL2007. UCREL. http://eprints.lancs.ac.uk/13011/1/192_Paper.pdf

Simon, Eszter (2014) *Corpus building from Old Hungarian codices.* In The evolution of functional left peripheries in Hungarian syntax. Oxford University Press, Oxford, pp. 224–36. ISBN 978-0-19-870985-5

Simon, Eszter & Veronika Vincze (2016) *Universal Morphology for Old Hungarian.* In Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@ACL 2016, 118–27. Association for Computational Linguistics.

Schneider, Peter (2002), *Computer assisted spelling normalization of 18th century English.* In P. Peters – P. Collins – A. Smith (eds.): New frontiers of corpus research: Papers from the 21st International Conference on English Language Research on Computerized Corpora, Sydney, 2000, 199–211. Rodopi, Amsterdam.

Vincze, Veronika; Szauter, Dóra; Almási, Attila; Móra, György; Alexin, Zoltán & Csirik, János (2010), *Hungarian Dependency Treebank.* In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 1855–62. European Language Resources Association (ELRA).