In our presentation we plan to discuss the various phases of a shortly terminating project,\(^1\) the aim of which was to create a morphologically annotated historical corpus representing the daily vernacular of speakers of Middle Hungarian. Therefore, the team selected samples of private correspondence and testimonies of witnesses in trials covering a period of three centuries from the end of the 15th century till the end of the 18th century. As it will be apparent from the examples introduced below, both the chosen period and these specific types of texts brought about quite challenging problems.

The essential first phase of the process was to compile a philologically adequate pool of digital texts. This was accomplished mainly through scanning, optical character recognition and thorough proofreading of the selected sources. As the texts were to be parsed by the Hungarian Morphological Analyzer HuMor developed for Modern Hungarian, the participants had to produce a normalized version of the sources, a transcription that could then be fed to the analyzer. This process comprised of the elimination of orthographical idiosyncrasies by modernizing to present-day orthography and neutralization of some dialectal phonological variation. In addition, the morphology implemented in the analyzer also had to be adapted to the task of annotating Middle Hungarian by extending the stem database and augmenting the affix inventory with items that have disappeared from the language but are present in our historical corpus. As a further consequence, a vast amount of transcription and annotation strategies emerged, all dedicated to enable correct morphosyntactic annotation on the one hand, and a correct representation of the grammar of the period on the other hand, including historical and partly dialectal morphological variation.

As problems concerning certain constructions/ambiguities and how these should be annotated became apparent during the process of annotating texts, convergence on specific final solutions was an iterative process. For instance, verbs were originally meant to be analysed as stem+inflection, for derivational morphology was thought to be beyond the scope of the project. (Verbal derivational morphemes were at the beginning considered to be part of the lemma.) However, it turned out that the HMM-based tagger that we used for morphological disambiguation performs significantly better if the analysis of certain derivational suffixes (e.g. the markers of the passive and the factitive) is represented in the morphosyntactic tags. Therefore the representation of the analysis of word forms containing these special derivational affixes was changed in the course of the project. Texts that had been annotated following the earlier method were all reanalyzed using automatic conversion of the analyses by regular expression patterns generated using the morphological analyzer. Similar reanalysis of already disambiguated texts was necessary at several points in the project whenever we were dissatisfied with some aspects of the annotations we had and decided to change it. Of course, the same changes usually also had to be applied to the morphological description implemented in the morphological analyzer.

With respect to the guidelines for the overall process of building the corpus, the main principle the team intended to follow throughout can be phrased quite simply: full conservation of morphological structure. In practice, however, a host of questions arose during normalization. One typical problem was the prevalence of ambiguous forms. Then again, ambiguity can have a lot of reasons, some of which needed unique solutions, others could be handled more generally. As various grammatical features of Middle Hungarian cannot be judged consistently on the basis of our Modern Hungarian competence, the team resolved the tagging conventions of the regularly ambiguous cases. For example, many instances of the word form \textit{az} were ambiguous, as it could be analyzed either as a deictic

\(^1\) Funded by the Hungarian Scientific Research Fund (grant nr. 81189).
pronoun or as a determiner. These cases were uniformly annotated as determiners, and the users will be reminded of this in the manual. Another practical solution to mark difficulties in analysis was to introduce special tags for annotation. A typical example of this is the third person form of the – now extinct – imperfective past, which could only be handled by introducing a special category label that codes the uncertainty of the analysis.

A further recurring issue was the assignment of elements into linguistic categories, with the corollary problem of the optimal level of categorization. If a category is too broad, it is easier to tag its putative elements, but the chance of getting problematic syntactic analyses on the clausal level grows significantly. If it is broken down into more fine-grained subcategories, the syntactic analyses will be more precise, but it complicates both the automatic tagging and the manual disambiguation. A straightforward example of this is the admittedly broad category of Adverbials. In this case, it seemed difficult to find the optimal level of elaboration. For instance, the annotation of a special subset, namely that of temporal adverbials, changed during the project, as it became evident that their syntactic distribution differs significantly from the rest of the adverbials. However, those adverbials that show conjunction-like features (by linking sentences), but lack the characteristic features of conjunctions (fixed position in the sentence) were annotated as adverbials.

In addition, adverbials pose problems from the point of historical morphology as well, which leads back to the problem of lacking (diachronic) competence. Many (if not all) of these arose through grammaticalization with the concomitant reanalysis of morphological structure, namely the loss of transparency. However, sometimes it is hard to decide whether a form was transparent in the given period (in which case the correct analysis would be to segment the form into a stem and a suffix). Moreover, from the point of a historical morphologist, synchronically unanalyzable forms can be invaluable as well. Finally, a new method was introduced to cover at least a significant subset of some problematic cases. Instead of substituting the already extinct adverbials with their functionally equivalent modern forms in the normalized texts, the lexicon of the parser was extended with the obsolete forms. Besides, we introduced the method of co-indexing the functionally equivalent adverbials in order to facilitate search in the corpus.

All in all, the participants of the project wished to build a corpus that is philologically reliable, yields correct analyses, and proves to be user-friendly for historical linguists and specialists or students of related fields as well. The corpus is freely available at the website of RIL, HAS, and all participants are welcome to use it and comment on it.

**References**

