The Parsed Corpus of Middle English Poetry
Richard Zimmermann

1. Definition and Motivation
The Parsed Corpus of Middle English Poetry (PCMEP) is a fully parsed and annotated corpus of Middle English verse texts. There is a substantial gap in English prose texts between c. 1250 and 1350. The main goal of the PCMEP is to help close this gap with available poetic records from the same period. In order to be able to assess the genre difference between prose and poetry, the corpus covers a slightly greater time span than that, namely c. 1150 to 1420 (Helsinki periods M1, M2, M3).

The PCMEP is parsed according to the same guidelines as its much larger sister corpus, the Penn-Parsed Corpus of Middle English, second edition (PPCME2). Thus, researchers familiar with the PPCME2 do not have to learn any new annotation schemes and can use their PPCME2 search queries without modification on the PCMEP text files as well.

2. Workflow
Appropriate texts are digitized from a large number of sources, followed by pre-processing, automatic parsing and finally manual correction.

3. Example of a Parsed Tree
The current text (May 2018) to be parsed is St. Margaret of Antioch (MEC *St.Marg.(2), IMEV 203, Manual 2.V.184(c)), composed c. 1220, 2,915 words. Horstmann (1881) serves as the base edition, out of copyright.

The parsed tree for the first sentence looks as follows:

4. Current Size and Goal
The corpus currently includes 42 Middle English poems with a total of 175,757 words. Detailed information on each text can be found on www.pcme.net. The ultimate goal for the corpus is to reach a size of 1 million words.

5. Example Application
Middle English replaces *then* with *when* as the subordinator in subordinate temporal clauses. (1)

Example 1: Conservative *pa* is replaced with innovative *when*

The change can be measured with great precision in the PCMEP. Statistical analyses return a rate of change which can then be compared to prose texts.

6. Conclusion
The PCMEP closes a substantial gap in the availability of parsed, digitized texts from the history of English. It allows fast, objective and accurate retrieval of linguistic data, and offers a new Middle English dataset to replicate previous findings and test new hypotheses.

Acknowledgments
The PCMEP was supported by a Doc.Mobility grant from the Swiss National Science Foundation (PIGEPE_148611).

References