University of Glasgow

12th International Conference on Middle English
22-24 August 2022

Index of Middle English prose: developing a search tool

Alpo Honkapohja, Jacob Thaisen (University of Oslo),

Finding medieval manuscripts is dependent on the search and reference tools available to us. The *Index of Middle English Prose* (IMEP), an expanding catalogue of opening and closing lines, is the most important reference tool for prose texts written in English between 1200 and 1500. More twenty printed volumes of the series have been published since the 1980s. Recently, IMEP is being digitised and a search tool capable of handling variation inherent in Middle English is under development. The aim of this paper is to discuss challenges related to it.

What makes developing an efficient search tool for Middle English difficult is the large amount of variation in a non-standardised vernacular. Variation is encountered in orthography, morphology, syntax and lexicon. A simple search for a string of words will find all occurrences of that specific string in a dataset and no other. Searches using regular expressions will also fail to meet the target since they presuppose that the user can predict every possible linguistic variation.

These challenges can be solved by developing a search tool based on natural language processing tools and methodologies, which are capable of coping with both the more straightforward orthographies and the more open-ended lexical and syntactic variation. It is possible to enhance the matching strings by using

similarity metrics based on edit distance or Jaccard distance, or by introducing a probabilistic element.

The data comes an ongoing project of completing an IMEP volume of the famous Cotton manuscripts located in the British Library. It is based on transcriptions made during fieldwork at the British Library, consisting altogether of 861 entries, each recording the incipits and explicits of an English prose text written between 1200 and 1500.