The Index of Middle English Prose

A search tool based on language modelling

Alpo Honkapohja, Jacob Thaisen and Anders Nøklestad
University of Oslo

The work has been funded by European Commission, Marie Skłodowska Curie: Individual Fellowship (MSCA-IF), Grant-ID 101025997.

What is *The Index of Middle English Prose* (IMEP)?

- The most important reference tool for Middle English non-verse texts composed between circa 1200 and 1500.
- There are currently 24 published volumes published by D.S. Brewer, Cambridge, with several more volumes in progress.
- An online version is available from Cambridge University Library (CUL)
- The new search tool for it is developed by the University of Oslo in collaboration with Cambridge Digital Humanities.

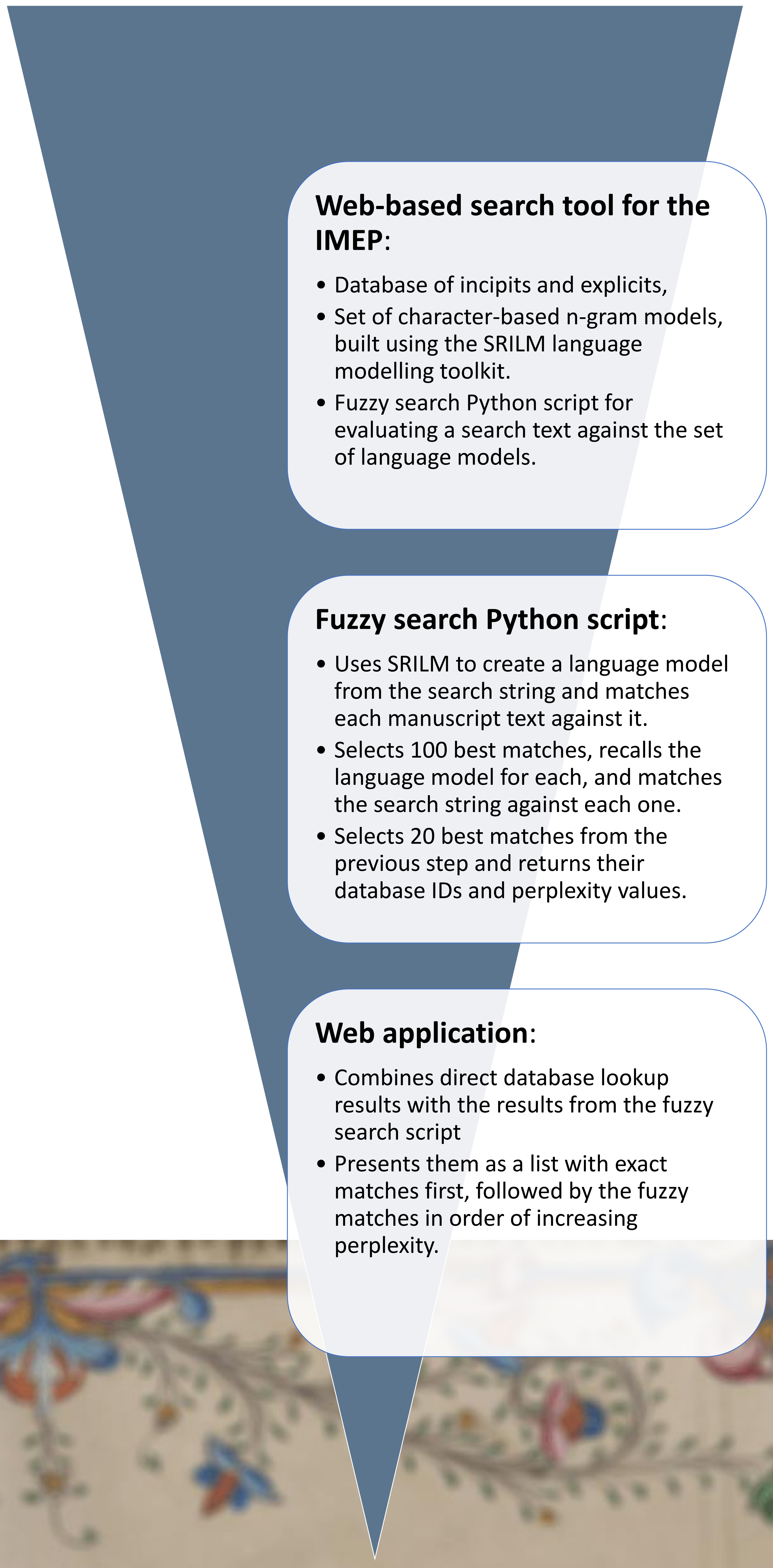
The problem:

Linguistic and textual variation in non-standardised early vernaculars:

Spelling, syntax, lexicon

Incipit	IMEP volume
In the nobele lande of syrrye there was a nobele kyng and	Rylands Eng 103 [1]
In the noble land of surey þer was a noble king	BL Add 10099 [1]
In the noble land of surre ther was a noble kyng and myhty and a	Lambeth 84 [1]
Off the noble land of surrye ther was a royal kynge	Peniarth 343 [1]
In the noble land of surrye ther was a worthi kynge	Lei UL 47 [1]
Some tyme in the lande of surre ther was a myghty & a ryall	Oxf Un 154 [1]

Our solution:



Web-based search tool for the IMEP:

- Database of incipits and explicits,
- Set of character-based n-gram models, built using the SRILM language modelling toolkit.
- Fuzzy search Python script for evaluating a search text against the set of language models.

Fuzzy search Python script:

- Uses SRILM to create a language model from the search string and matches each manuscript text against it.
- Selects 100 best matches, recalls the language model for each, and matches the search string against each one.
- Selects 20 best matches from the previous step and returns their database IDs and perplexity values.

Web application:

- Combines direct database lookup results with the results from the fuzzy search script
- Presents them as a list with exact matches first, followed by the fuzzy matches in order of increasing perplexity.

UNIVERSITY OF CAMBRIDGE
Cambridge University Libraries