Creating an extended author's dictionary to support digital literary research

Margit Kiss
Institute for Literary Studies,
Research Centre for the Humanities,
Hungarian Academy of Sciences
kiss.margit@btk.mta.hu

Tamás Mészáros
Dept. of Measurement and Information Systems,
Budapest University of Technology and Economics
meszaros@mit.bme.hu

Storing literary and historical text in digital form and using software tools to process and analyze them give researchers potentials to perform new kinds of research tasks and to achieve new scientific results. A significant problem, however, is that it is hard to cope with the diversity and variations of the grammar and vocabulary of texts from the distant past. Statistical methods relying on word frequency counting and text clustering fail to produce accurate results when the same word takes many different forms. Searching also poses problems to researchers as it is hard to keep in mind all kind of word alterations. Text normalization techniques based on probabilistic learning models are used to overcome these problems (Oravecz 2010) but in many cases they fail to produce high quality results. We faced these issues while investigating a large text corpora (roughly 1,5 million words) from the XVIII century: the works of Kelemen Mikes often called the “Hungarian Goethe”.

Our approach of solving these problem was to create an extended author's dictionary that can help in developing better analysis and search tools. Elaborating an author's oeuvre in dictionary form is a known technique in literary research (Karpova 2011). The purpose of a simple concordance is to support the researchers orientation and search in the corpora. We extended this concordance list with additional lexicographic categories and with linguistic annotations. Our method can be seen in a way that it gathers information from close reading methods and it makes them available for distant reading techniques in order to increase their effectiveness.

We developed a software application that analyzed the corpora and automatically created the initial version of the dictionary including all word forms, sample sentences and their references. Then this was manually reviewed

and edited: the alternative forms and writing variations of the same dictionary word were unified into a single dictionary entry. During this editing phase entries were also extended with modern dictionary words, reference headwords and annotations marking proper names, author's word creations, words with foreign language origins, etc. The final result contains more than 20000 dictionary words and 162000 word forms (Kiss 2012). It provides far more information than a simple concordance list of words and their occurrences and it can help researchers to perform new kinds of analysis of the corpora to uncover grammatical, literary, rhetorical and other attributes and results.

To present the information to researchers we also created a Web application that displays the XML data set in human-readable form. This tool also facilitates search using both original and modern word forms therefore we could avoid the problems of morphological analysis of historical texts. While browsing the dictionary researchers can also read text excerpts and follow their links to the original works of Kelemen Mikes.

There are many potential applications of this extended digital author's dictionary. Firstly, researchers can analyze the dictionary data itself using programmable tools. These may include computing statistics about words with certain kinds of annotations (like the author's own word creations, proper names or words with foreign origins). It may also be possible to compare various author's dictionaries and analyze vocabulary similarities and influences or changes over time. Since this extended dictionary organizes the corpora in a very different way its analysis may also uncover previously hidden connections or similarities between distant text pieces.

Secondly, this data set could enhance the quality of full text analysis and search services. For example, search terms and indexes can be normalized by replacing text variations and old word forms with their modern variant found in the dictionary. This way searching could be done using any kind of word forms and its precision could be greatly enhanced. In order to demonstrate the benefits we created a simple comparison with traditional full text search. The traditional search only provided a few results for the word “Constantinople” (which has many writing variations). Our dictionary-based normalized search provided 105 results which is a great increase in recall. We also performed a full-text manual search to validate the results.

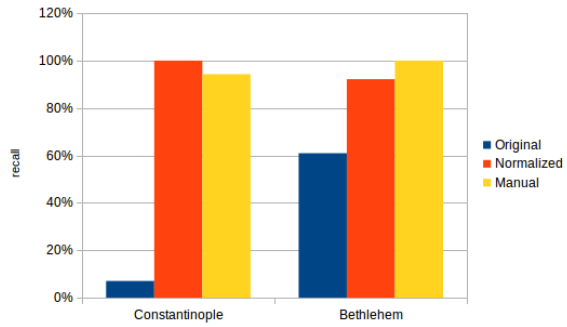

Figure 1: Performance evaluation of dictionary-based search

Statistical analysis (e.g. authorship attribution) tools can also be extended with a preprocessing (text normalization) phase using this dictionary. Certain word forms could be replaced with their standard variants in order to increase the quality of the analysis. These text analysis methods could also incorporate information found in dictionary annotations. For example, an analysis could be performed relating the origins of the words (own creations, loan words, etc.) found in the corpora.

Lastly, the author's work could be annotated using the information found in the dictionary. This includes the classical use of an author's dictionary: helping readers in understanding words and phrases, but it may include other services based on annotations or text analysis results. These information can be presented as notes to the user reading the original works and they also could be a basis for creating detailed critical editions.

The digital author's dictionary greatly increases the quality of software tools analyzing literary or historical texts from the distant past and it may also open up new opportunities for studies like examining special features of language usage, style and the changes of these, and to answer more complex questions relating to the author’s view of life. Future work may include incorporate meanings of dictionary words and describing semantic relations between them using ontology tools. Dictionary entries and annotations may also be linked to external data sources like WordNet, Wikidata or Dbpedia using Linked Open Data techniques.

References

Oravecz, Cs., Sass, B. and Simon, E. (2010) 'Semi-automatic normalization of Old Hungarian codices', *Proceedings of the ECAI Workshop on Language Technology for Cultural Heritage,* pp. 55–59.

Karpova, O. (2011) *English Author Dictionaries*, Newcastle upon Tyne: Cambridge Scholars Publishing.

Kiss, M. (2012) 'The Digital Mikes-Dictionary', Tüskés Gábor, Bernard Adams, Thierry Fouilleul, Klaus Haberkamm (eds.) *Transmission of Literature and Intercultural Discourse in Exile,* Bern: Peter Lang Verlag, pp. 288-297.