Digital thesauri as semantic treasure troves: a Linguistic Linked Data approach to "A Thesaurus of Old English"

Stolk, S.S.

Conclusion

This thesis comprises an investigation into how Web-based dissemination of historical language thesauri can be improved so as to answer to the research needs of scholars in various disciplines. A synopsis of the most noteworthy results is provided below. Next, this conclusion discusses areas for future research and, lastly, it considers the original contributions to onomasiological studies and lexicography made by this dissertation.

The investigation started by reviewing characteristics of historical language thesauri. Chapter 1 offered insights into the content of these lexicographic resources. A thesaurus – i.e., a semantically organized dictionary – consists of three main parts: (1) the topical system, which is a hierarchy of semantic concepts; (2) lexical senses, which are words or phrases in a specific sense, positioned within the overarching topical system; and, optionally, (3) relations of synonymy, indicated through groupings of lexical senses. The information captured for each of these parts varies from thesaurus to thesaurus. Which usage features of lexical senses are included, for instance, is largely determined by the intended use of the thesaurus. Nevertheless, as the overview in Chapter 2 indicated, researchers want to use the thesaurus content as a stepping stone for investigating related matter outside the scope of the originally intended purpose of the resource. The ability to reuse and elaborate on thesaurus content and to subsequently perform onomasiological analyses on the newly available, combined knowledge constitutes the most notable, novel pieces of functionality desired for research that are currently lacking in the majority of the published historical language thesauri of Scots and of English.

Following the analysis of the content and functionality desired of historical language thesauri, Chapters 3 to 5 considered in what digital form these lexicographic resources should be published on the Web. The digital form proposed is one based on Semantic Web standards, most notably SKOS and Lemon-OntoLex, and has more recently been termed Linguistic Linked Data. The current specifications of these two Semantic Web standards proved insufficient for representing thesauri fully; my work on the development of the *lemon-tree* model was intended to address this matter. In addition to stipulating the manner in which the two data vocabularies should be combined in order to capture the content of thesauri, *lemon-tree* covers two aspects important for representing these resources: (1) levels that can be distinguished in the topical system of these works and (2) a looser form of categorization than lexicalization. These additions are relevant not only for historical language thesauri, but for thesauri in general.

The last four chapters evaluated the usefulness of the Linguistic Linked Data form for thesauri by applying it to *A Thesaurus of Old English* (*TOE*), a historical language thesaurus that captures the early medieval English lexicon. The result of the transformation of *TOE* from its original database format to Linguistic Linked Data (or *TOE*-LLD), detailed in Chapter 6, was disseminated through the web application Evoke, tailored to thesauri and newly developed as part of this dissertation. Evoke and *TOE*-LLD were utilized in various case studies within the collaborative research project 'Exploring Early Medieval English Eloquence' (EEMEE). Chapter 8 offered an overview of these studies, their approaches and results, and reflected on their use of the thesaurus and the web application. The assessment of the digital form and the functionality offered by Evoke, in terms of their usefulness for research and education, foregrounded that their combination forms a powerful tool for researchers to enrich original thesaurus content with additional data. Researchers have demonstrated they could use Evoke to perform novel research on such topics as lexicographic history, stylistics, diachronic developments, and kindred languages. Onomasiological analyses possible through Evoke, which were used on combined sets of information, have led to new insights into Old English language and culture for both students and researchers. Published articles on the case studies underline the value of these new tools available for research and, more often than not, propose future areas of research that should prove fruitful.[1]

One avenue for future work is to continue investigations into Old English language and culture, utilizing Evoke and *TOE*-LLD, by extending the original thesaurus content with additional data. As Thijs Porck and Amos van Baalen have demonstrated, Evoke and *TOE*-LLD facilitate the creation of textual or authorial subthesauri.[2] Their approaches yielded the *Beowulf Thesaurus* and *Ælfrician Vocabulary*, amongst others, and established onomasiological profiles that capture the lexical choices and preferences within a text or of a specific author. Similar investigations of Old English texts, authors, and genres beyond those currently covered can, as Porck suggests, "form the starting point of new inquiries into these age-old texts".[3] Furthermore, onomasiological profiles would benefit from further refinement. Incorporating attestations through integrating links to digital corpora and ascertaining preferences of certain words over not just their synonyms, but also over available hypernyms and hyponyms, are two examples of alterations that would result in finer-grained semantic fingerprints.

In addition to further work with *TOE*, future research can branch out to other thesauri as points of departure. Other historical languages, contemporary languages, and sign languages may well be equally suitable for research. Some of these explorations will benefit from the reuse of an existing topical system, which Rita van de Poel and Sander Stolk have shown to be advantageous for comparative analyses;[4] for others, a semantic macrostructure may already be in place for the lexis concerned or demand building one from the ground up.

[1] See Chapter 8 for an overview of these case studies and references to the various journal articles that describe them in detail.

[2] Ibid.

[3] Porck, 'Onomasiological Profiles of Old English Texts', p. 379.

[4] Van de Poel and Stolk, 'A Case of Kinship'. The article is included in full as Chapter 9 of this dissertation.

As the EEMEE project has demonstrated, subsequent analyses and extension of thesauri through a collaborative research programme can pave the way for explorations and discussions. Workshops facilitated the sharing of insights on the thesaurus, additional information gathered and connected to the lexicographic resource, and the functionality available for researchers and students to interact and query the material at hand. The engaging environment propelled efforts undertaken for this dissertation and aligned further development of the web application Evoke with the needs of those who used it. Even so, participants in the project indicated they experienced hurdles in connecting additional data to an existing thesaurus, which proved to be time-consuming for larger sets of information when done manually,[5] imprecise when using (semi-)automatic linking strategies,[6] and challenging due to differences between the thesaurus and additional data in terms of the inherent lexicographic choices and mismatching conceptualisations as represented in the onomasiological framework.[7] Challenges such as these are not limited to linking data to historical language thesauri, but are common themes in resource alignment.[8]. Although there are no easy solutions, offering insights into these matters alongside digital thesaurus editions – perhaps in the form of a reference guide that includes ways researchers have overcome (or circumvented) such difficulties – would remove some of the hurdles in extending and using historical language thesauri.

Experiences at University of Groningen and Leiden University are encouraging for use of Evoke and *TOE*-LLD for educational purposes. Both universities developed their own set of assignments for students and incorporated the material into their Old English curricula. Kees Dekker posits that teachers in similar contexts would benefit from existing material to be made publicly available alongside the two digital resources.[9] In order to answer this call, a set of questions and assignments for the Old English classroom are planned to be made available on the Evoke website. By courtesy of Prof. Carole Hough (University of Glasgow), the exercises will include units from the module *Learning with the Online Thesaurus of Old English*. Preliminary work has been done towards incorporating these assignments. Since figures on the degree of lexicalization of categories and semantic domains appear prominently in them, these learning modules will benefit from the novel functionality of Evoke to offer automated analyses on this key aspect.

In all, the work in this dissertation has centred on the use of Linguistic Linked Data as the digital form of thesauri for research. Proclaimed benefits of this form

[5] Both Porck and van Baalen called on a student-assistant to perform the bulk of the alignment (see Porck, ‘Onomasiological Profiles’, p. 381; van Baalen, ‘Identifying, Categorising and Exploring “Ælfrician” Vocabulary’, p. 408).

[6] Depuydt and de Does, ‘Linking the *Dictionary of Old Dutch* to *A Thesaurus of Old English*’.

[7] E.g., Porck, ‘Onomasiological Profiles’, pp. 364-9; Van de Poel and Stolk, ‘A Case of Kinship’, pp. 477-8.

[8] An example from a lexicographic context is word sense alignment. See, for instance, Ahmadi and McCrae, ‘Monolingual Word Sense Alignment as a Classification Problem’. Moreover, tooling is actively being developed for aligning of Linked Data resources automatically. Examples are Amalgame (https://github.com/jrvosse/amalgame) and NAISC (https://github.com/insight-centre/naisc)

[9] Dekker, ‘Evoke and *A Thesaurus of Old English* in the Old English Classroom’, p. 526.

put forward by Christian Chiarcos et al. – merging of datasets, interoperability, linking data rather than duplicating – were consistent with the functionality required for research as catalogued in Chapter 2.[10] These benefits have indeed facilitated researchers to reference and extend thesaurus content and can be recommended on those grounds for use in research that demands such features. Still, a comprehensive analysis that contrasts various digital forms for these (and other) case studies has not yet been undertaken. Other digital forms (such as SQL or, document-based, XML and JSON) might prove to be more effective in certain areas. A comparison of this kind could take into account such factors as the effort and time required to implement functionality on top of a given digital form for thesauri, which depends, amongst others, on the availability of databases for these formats and of software libraries that can be reused to bootstrap development of new applications. Similarly, the costs of hosting an application built around a digital form and responsiveness of the functionality offered are two other factors worth contrasting between possible solutions.[11] As indicated in the introduction of Chapter 3 of this thesis, "Each form has its strengths and weaknesses, making some forms better suited for a specific purpose than others". As a consequence, a solution that fits all needs is unlikely to exist. Which digital form is most appropriate may best be decided on a case-to-case basis.

With respect to onomasiological studies, this thesis has made contributions in the form of new statistical analyses that have been developed. These analyses provide insight into the distribution of lexis in a thesaurus and allow for salient features to be selected on which to focus specifically. The analyses currently available in Evoke include the degree of ambiguity of selected lexical items, their degree of synonymy, and their distribution over the onomasiological framework of the thesaurus in terms of its categories and the depth of the taxonomy.[12] The first-mentioned analysis, which is based on polysemy, is useful for signalling the tendency of authors to either employ or avoid ambiguous words in their writings. The second, the degree of synonymy, can be used to convey the number of alternatives, or choice set, available to authors in making their lexical choices.[13] The two analyses on the distribution communicate the dominance of semantic domains and the level of precision in the meanings attributed to the selection of lexical items. The analyses contribute towards a semantic fingerprint, or onomasiological profile, that maps characteristics of lexis, which can facilitate onomasiological comparisons and prompt new inquiries based on these findings. Although these profiles are promising, further study is required to ascertain their most desired constituents, the contexts in which specific analyses are

[10] Chiarcos et al., 'Towards Open Data for Linguistics: Lexical Linked Data'.

[11] See, for instance, the comparison between relational databases and graph databases for specific benchmark datasets by Cheng et al., 'Which Category Is Better', which focuses on query response time, but also offers an indication of CPU and memory usage for query processing. Figures such as those presented in this article are informative for the amount of server-side resources expected to be required, which impacts hosting costs.

[12] Stolk, 'Evoke', p. 342.

[13] Note that the choice set of authors may be limited due to constraints imposed by the context, such as the metrics or stylistic devices demanded in poetry.

merited, and the conclusions that can be drawn safely from their composition. As mentioned in Chapter 8, a thesaurus presents a filtered image of the lexicon it captures (e.g., bias through its corpus, existing interpretations, lexicographic choices).[14] For optimal use, therefore, that image – and outcomes of analyses drawing on that image – should be fully understood.

Next to onomasiological studies, this thesis has made a number of contributions towards lexicography and related Semantic Web standards. Active participation in the W3C Ontology-Lexicon (OntoLex) Community Group has resulted in the publication of additional modules for the Lemon-OntoLex standard in representing Linguistic Linked Data: 'The OntoLex Lemon Lexicography Module' for representing dictionary structures, *lemon-tree* for thesauri, and, forthcoming, a module for modelling frequency and attestations of lexis in corpora.[15] These technical specifications and corresponding data vocabularies have enabled important facets of lexicography to be expressed as Linguistic Linked Data. Their availability should assist in furthering areas in which this digital form can be applied — both in and outside of research contexts.

Other major contributions this thesis has made to the field of lexicography are constituted by novel web applications and data transformations. The most notable amongst these is Evoke. This application, developed over four years, demonstrates how sets of functionality desired by researchers catalogued in Chapter 2 can be implemented in a user interface. The source code of the latest version, 1.4.1, is publicly available on GitHub under the GPL 3.0 license.[16] As Chapters 8 and 9 have shown, Evoke has already been used effectively in the scholarly world for research and education. Publishers of thesauri and other lexicographic resources can assess the usefulness of the features implemented here and reuse, or draw inspiration from, the source code for incorporating those features they deem valuable for their audiences. Noteworthy elements of its design are the use of IRIs in order to facilitate data links, an annotation system that allows users to capture their contributions explicitly, and analysis functions to scrutinize the onomasiological distribution of lexical items of interest.

Besides Evoke, this dissertation includes source code for both data transformations and alignment tools.[17] The data transformations include that of *TOE* to its Linguistic Linked Data form, *TOE*-LLD, as discussed in Chapter 6, and the transformation of the 'Old Frisian: Kinship' dataset from Microsoft Excel to Linguistic Linked Data, which was used in the article positioned as Chapter 9. The alignment tools that have been made readily available to the public, too, are those used in the EEMEE case study by Thijs Porck for aligning lexis found in glossaries of Old English texts with that recorded in *TOE*-LLD.[18] The public availability of these transformations and tools illustrates a method of

[14] See section 8.2.

[15] See 'The OntoLex Lemon Lexicography Module'; Chapter 5 on *lemon-tree*; and Chiarcos et al., 'Modelling Frequency and Attestations for OntoLex-Lemon'. The homepage of the W3C Ontology-Lexicon Community Group is https://www.w3.org/community/ontolex/.

[16] See 'List of source code' in the back matter of the dissertation.

[17] Ibid.

[18] Porck, 'Onomasiological Profiles of Old English Texts'. A short description of the case study is provided in Chapter 8.

expressing thesauri as Linguistic Linked Data and exemplifies how links between datasets can be captured and shared in that same format.

Beyond the insight into thesauri and beyond the developed web applications and automated data transformations, work on the dissertation contributed to academic research by bringing together a group of researchers to explore various facets of Old English and related historical languages. The EEMEE research programme has resulted in a number of case studies on the application of a historical language thesaurus (i.e., *TOE*) in a newly available, digital setting (i.e., Evoke). Their findings, of which many have been published in a special issue of the international, peer-reviewed journal *Amsterdamer Beiträge zur älteren Germanistik* (volume 81, issues 3-4), are as various as the disciplines involved. Amongst these findings are onomasiological characteristics of the lexis in Old English texts (i.e., *Beowulf*, *Andreas*, the *Old English Martyrology*) and that employed by Ælfric of Eynsham, specifically; differences between the kindred languages Old Frisian and Old English in the semantic field of KINSHIP; a catalogue of Old English metaphors for SHAME and their conceptual mappings; and findings of tutors and students working with a historical thesaurus. Many of the datasets fashioned in these case studies are now publicly available, too, allowing anyone to explore and analyse the *Beowulf Thesaurus*, or any of the other sets of information, alongside *TOE*.[19]

Lastly, the wealth of resources made available for research, listed above, alludes to the impact and potential of the work put forward in this dissertation. Their availability, open to all, is intended to stimulate collaboration and creativity of researchers and to encourage further ventures into Digital Humanities by scholars and students alike. The dissertation (and its author) are indebted to the community of philologists, linguists, historians, lexicographers, and students who provided input, used, or engaged otherwise with these materials. The results of their combined efforts have strengthened the notion that, for use in academia, there is yet ample opportunity to improve the Web-based dissemination of historical language thesauri. The facets explored in this dissertation, and with others still unexplored, demonstrate that thesauri are by no means exhausted by previous investigations. Digital thesauri are still – in new and exciting ways – semantic treasure troves.

References

Ahmadi, S. and J. P. McCrae, ‘Monolingual Word Sense Alignment as a Classification Problem’, Proceedings of the 11th Global Wordnet Conference, 18-21 January 2021, pp. 73–80. https://aclanthology.org/2021.gwc-1.9.pdf.

Cheng, Y. et al., ‘Which Category Is Better: Benchmarking Relational and Graph Database Management Systems’, *Data Science and Engineering* 4 (2019), 309–22. doi: 10.1007/s41019-019-00110-3.

Chiarcos, C. et al., ‘Towards Open Data for Linguistics: Lexical Linked Data’,

[19] See Appendix 8.A.

in *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, eds. A. Oltramari, P. Vossen, L. Qin, and E. Hovy (Heidelberg, 2013), pp. 7–25.

Chiarcos, C. et al., 'Modelling Frequency and Attestations for OntoLex-Lemon', Proceedings of the 2020 Globalex Workshop on Linked Lexicography, Marseille, 20 June 2020, pp. 1–9. https://aclanthology.org/2020.globalex-1.1.

Dekker, K., 'Evoke and *A Thesaurus of Old English* in the Old English classroom', *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), 514–29. doi: 10.1163/18756719-12340241.

Depuydt, K. and J. de Does, 'Linking the *Dictionary of Old Dutch* to *A Thesaurus of Old English*: A First Exploration', *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), 493–513. doi: 10.1163/18756719-12340240.

'Lexicon Model for Ontologies', eds. P. Cimiano et al., *W3C*. http://www.w3.org/2016/05/ontolex/. Final Community Group Report. Created: 10 May 2016.

Porck, T., 'Onomasiological Profiles of Old English Texts: Analysing the Vocabulary of *Beowulf*, *Andreas* and the *Old English Martyrology* through Linguistic Linked Data', *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), 359–83. doi: 10.1163/18756719-12340236.

'SKOS Simple Knowledge Organization System Reference', eds. A. Miles and S. Bechhofer, *W3C*. http://www.w3.org/TR/skos-reference/. W3C Recommendation. Created: 18 August 2009.

Stolk, S., '*lemon-tree*: Representing Topical Thesauri on the Semantic Web', Proceedings of the 2nd Conference on Language, Data and Knowledge (LDK 2019), Leipzig, 20-23 May 2019. doi: 10.4230/OASIcs.LDK.2019.16.

Stolk, S., 'Evoke: Exploring and Extending *A Thesaurus of Old English* using a Linked Data Approach'. *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), pp. 318-58. doi: 10.1163/18756719-12340235.

'The OntoLex Lemon Lexicography Module', eds. J. Bosque-Gil and J. Gracia, *W3C*. https://www.w3.org/2019/09/lexicog/. Final Community Group Report. Created: 17 September 2019.

Van Baalen, A., 'Identifying, Categorising and Exploring "Ælfrician" Vocabulary using the *Dictionary of Old English*, *A Thesaurus of Old English* and Evoke', *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), 384–441. doi: 10.1163/18756719-12340237.

Van de Poel, R. and S. Stolk, 'A Case of Kinship: Onomasiological Explorations of KINSHIP in Old Frisian and Old English', *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), 457–92. doi: 10.1163/18756719-12340239.