Universiteit
Leiden
The Netherlands

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

Stolk, S.S.

Citation
Stolk, S. S. (2023, May 31). *Digital thesauri as semantic treasure troves: a Linguistic Linked Data approach to "A Thesaurus of Old English*". Retrieved from https://hdl.handle.net/1887/3619351

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3619351

Note: To cite this publication please use the final published version (if applicable).

Digital Thesauri as Semantic Treasure Troves

© 2023 Sander Stolk

Lay-out by Sander Stolk
Cover design by Sander Stolk & Martin Stolk

Digital Thesauri as Semantic Treasure Troves

A LINGUISTIC LINKED DATA APPROACH TO *A THESAURUS OF OLD ENGLISH*

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 31 mei 2023
klokke 10.00 uur

door

Sander Sebastiaan Stolk

geboren te Beverwijk
in 1985

Promotores:	Prof. dr. R. H. Bremmer Jr
	Prof. dr. P. Th. J. M. Vossen (VU University, Amsterdam)
Copromotor:	Dr. M. H. Porck

Promotiecommissie:	Prof. dr. M. Alexander (University of Glasgow)
	Prof. dr. J. Grijzenhout
	Prof. dr. L. C. J. Barbiers
	Dr. M. Cole (Utrecht University)

Preface

After six years of working as an external PhD researcher, two days a week, I finished the dissertation in front of you. Doing research next to a demanding job has by no means been an easy feat, not to mention the number of academic activities I engaged in beyond those strictly necessary for the dissertation. Looking back, I do not recommend others to employ a similar strategy and increase the workload so, however fulfilling the additions may be. Even so, I am thankful for the opportunities each activity presented me and, first and foremost, the strong professional and personal connections that resulted from them (to which this preface testifies).

I would like to express my gratitude to my supervisors Rolf Bremmer, Thijs Porck, and Piek Vossen for their insights and suggestions. Thijs, I hold dear your support, especially. We have co-organized two workshops and co-edited a special issue journal. Your words and encouragements helped me progress and ensured I celebrated key milestones. I would also like to thank the team behind *A Thesaurus of Old English* wholeheartedly; Jane Roberts, Marc Alexander, and Fraser Dallachy in particular. You were kind enough to allow me to use this treasure trove in my research and inspired me to explore its potential applications — an exploration for which I am greatly indebted to participants in the research programme Exploring Early Medieval English Eloquence. Your case studies, suggestions, and feedback have been of utmost importance in developing the resources at the heart of this thesis.

Further communities enriched my academic life: the W3C OntoLex Community Group, closely collaborating on standardizing Linguistic Linked Data; the Leiden University Centre for the Arts in Society (LUCAS), at which I performed my research and which granted me a fund to finish this dissertation; the Leiden University Centre for Digital Humanities (LUCDH), which offered advice and supported me through small grants; and the Digital Humanities Student Network (DH/SN), in which I found myself amidst peers with a shared passion. Thank you for welcoming me as part of your respective communities.

Additionally, I would like to thank my colleagues at Semmtech, my bouldering buddies, my friends and family for your support and, most of all, your company. Lastly, I would like to thank my girlfriend-turned-fiancée-turned-wife, Fenja Schulz. Our journey together has been one of kindness, understanding, and growth. May we continue celebrating what we have and what we achieve, both together and as individuals.

Acknowledgements

The dissertation was supported by the LUCAS Extra Resources Open Call-II Grant (5.000 EUR) for the research project Exploring Medieval English Eloquence and by the Afrondingssubsidie buitenpromovendi (18.000 EUR), awarded by the Leiden University Centre for the Arts in Society (2020; 2022). The early development of the web application Evoke was supported by the LUCDH Small Grant (3.500 EUR), awarded by the Leiden University Centre for Digital Humanities (2018).

Contents

Introduction

"*Change*, *alter*, or perhaps *transform*?" Selecting the perfect word for a specific context, such as when composing a report or a speech, is all the easier with a thesaurus at hand. These lexicographic resources are invaluable for looking up alternative words or phrases that convey a specific meaning. Indeed, thesauri are sometimes included as course material to acquire a language as well as the nuances available within one.[1] In addition, thesauri offer a number of uses beyond looking up alternative phrasings: they are veritable treasure troves for cultural, linguistic, anthropological, and literary-critical research — especially when these resources are arranged in a topical fashion, a hierarchical ordering of its groups of loosely synonymous words according to their meaning.[2] These topical thesauri – or, more specifically, ones that cover a historical language – take centre stage in this dissertation, which examines how their dissemination on the Web can be improved to facilitate research.

This introduction is organized as follows. Section 0.1 introduces the main topic of my dissertation, i.e., thesauri, more specifically, thesauri of historical languages. Next, section 0.2 offers an overview of the various research applications of a thesaurus. Section 0.3 discusses opportunities for improvement of the dissemination of these resources to this end. In section 0.4, the research objective and questions can be found, followed by an outline of the dissertation in section 0.5. Lastly, section 0.6 provides an overview of related work.

0.1. Thesauri and historical languages

Historically, the term *thesaurus* has been applied to a range of resources. The term, derived from Greek *thēsauros* ("store, treasure"),[3] has been used in the sense of a 'classical lexicon', a 'semantically organized dictionary', and a 'terminological database' or 'index'.[4] The first sense is now largely obsolete;[5] the

[1] See, for instance, Gánem-Gutiérrez and Gilmore, 'A Mixed Methods Case Study on the Use and Impact of Web-based Lexicographic Tools on L2 Writing'. The authors state that thesauri, and other Web-based lexicographic resources, "have become widely used by second language (L2) learners, particularly for academic writing" (p. 1). Celen and Yalçin demonstrate benefits of such use in their article 'The Effects of Vocabulary Resource Use on Lexical Richness in L2 Writing'. Turkish students' consulting of a thesaurus "led to a significant increase in lexical sophistication" in their English essays (p. 1039).

[2] Brewer, Review of *Historical Thesaurus of the Oxford English Dictionary*, p. 802; Adamska-Salaciak, Review of *Historical Thesaurus of the Oxford English Dictionary*, p. 232; Busse, 'A Celebration of Words and Ideas', p. 88.

[3] *Oxford English Dictionary*, 2nd edn, s.v. 'thesaurus, n.'.

[4] Hartmann, *Encyclopedia of Language & Linguistics*, s.v. 'Thesauruses'

[5] An example of a thesaurus in the sense of a classical lexicon is the *Thesaurus Linguae Latinae*, an authoritative dictionary of ancient Latin that, according to the introduction provided on

second and third, developed through semantic narrowing, are still current and, as will become apparent in the dissertation, are closely related to one another.[6] The dissertation concentrates on the second meaning of the term,[7] which has been defined more specifically as "a work of lexicographical reference which presents lexical facts with semantic domains as its core organizational principle, rather than in alphabetical arrangement".[8]

The first semantically organized dictionary, published in 1852, was Peter Mark Roget's *Thesaurus of English Words and Phrases* (henceforth *Roget's*). This work by Roget (1779-1869), a British physician and theologian, was the first of its kind to arrange a lexicon – that of contemporary English – in groups of loosely synonymous words according to an overarching, hierarchical structure.[9] This macrostructure can be likened to the taxonomies of animals and plants by Carl Linnaeus (1707-1778). In fact, Roget may even have taken these taxonomies of the natural kingdom as examples for his own structure.[10] In these tree-like structures, the most generic or abstract concepts are used as roots, which branch out to groups of words increasingly specific in meaning.[11] From the start, *Roget's* was commercially successful and has remained popular to the present day — with a distribution "almost comparable to that of the Bible".[12]

Since the publication of *Roget's*, a small number of thesauri have been fashioned that capture the lexicon belonging to a historical period rather than a contemporary one.[13] Amongst these (though not the first) is *A Thesaurus of Old English* (*TOE*). *TOE*, first published in 1995, is concerned with the Old English lexicon, the variant of English spoken between roughly 500 and 1100 by the Anglo-Saxons. This thesaurus has been met with high praise by scholars. Rolf H. Bremmer Jr, for instance, states that the thesaurus fills a "voluminous gap [...] on the shelf of lexicographical tools" available for Old English.[14] Richard Dance, too, calls *TOE* "invaluable" for lexical studies and deems it an "impressive piece of scholarship".[15] Manfred Görlach goes so far as to state that *TOE* is "the most important contribution to Old English studies for years", since its "comprehensive analysis" allows scholars to "investigate what distinctions Anglo-Saxons felt important enough to make in the lexicon".[16] Like *Roget's*, *TOE* treats its lexicon as a temporally indistinguishable whole. Such

its website, covers "all surviving Latin texts from the earliest times down to AD 600".

[6] See Chapter 3.

[7] In articles that demand distinguishing between the second and third sense, I apply the terms *topical thesaurus* and *indexing thesaurus*, respectively. Examples of the latter are the 'NASA Thesaurus', 'EuroVoc', and the 'Medical Subject Headings RDF'.

[8] Kay and Alexander, 'Diachronic and Synchronic Thesauruses', p. 367.

[9] Hüllen, *A History of Roget's* Thesaurus, p. 234.

[10] Ibid., p. 18.

[11] Onomasiological approaches to language, which adopt a thematic arrangement of words and phrases rather than an alphabetical one, have a long history predating *Roget's*. Early works of this kind, albeit not capturing the lexicon of an entire language, date back as far as Antiquity and possibly further still (Hüllen, *A History of Roget's* Thesaurus, p. 44).

[12] Ibid., p. 1.

[13] See Chapter 1.

[14] Bremmer, 'Treasure Digging', p. 109.

[15] Dance, Review of *A Thesaurus of Old English*, p. 312.

[16] Görlach, Review of *A Thesaurus of Old English*, pp. 398-9.

thesauri are called synchronic. A diachronic one, charting the changes throughout a certain period, did not yet exist of an entire language when *TOE* was published. However, *TOE* was intended as a pilot of such a thesaurus.[17]

The first diachronic thesaurus of an entire language was published in 2009: *The Historical Thesaurus of English* (*HTE*).[18] This thesaurus charts the development of the entire English lexicon, from the Old English period up to the present. Its information on the Old English period came from *TOE*; information from later periods was taken from the *Oxford English Dictionary*. The annotation added for lexical items on period of use allows these items to be ordered not just by their meaning but also chronologically.[19] This diachronic approach for the English language as a whole allows for comprehensive investigations into semantic change.[20] Additionally, the thesaurus allows for a focus on the vocabulary available in a specific time frame, such as that available to Shakespeare,[21] providing the opportunity for more thorough investigations of historical stylistics. Owing to the new paths opened up to them, researchers have dubbed the thesaurus "invaluable" or even a "godsend", underlining the importance of such semantically organized dictionaries.[22]

0.2. Thesauri and their applications

In addition to providing the means to locate and select available alternative phrasings, thesauri offer a number of applications valuable for research into language and culture. To illustrate, thesauri facilitate analyses and comparisons of semantic fields: groups of words related in meaning. Since lexical items located near each other in a topical thesaurus are by definition related in meaning, the words that make up a semantic field (e.g., for emotions) can more easily be pinpointed and subsequently scrutinized.[23] Moreover, when a thesaurus contains indications of the time frames in which words were used for specific meanings, the thesaurus can be used to research the development of the language it deals with.[24] To give an example, the word *knight* has gained a meaning during the Middle English period that is more positive than in the preceding period: it

[17] Roberts, '*A Thesaurus of Old English*: The Pilot Study for the Glasgow *Historical Thesaurus*'.

[18] Information on its versions can be found in the section 'Versions of the Thesaurus'. The first version of this thesaurus was published as *Historical Thesaurus of the Oxford English Dictionary*, and its contents are still available digitally on the website of the *Oxford English Dictionary*. Although this thesaurus is the first diachronic one of an entire language, thesauri acting as pilots and foretastes have been published at earlier points in time. See *TOE* and Coleman, *Love, Sex and Marriage*.

[19] The dissertation adopts the term *lexical item* to refer to a word or phrase in a single sense or across all of its senses. As such, *lexical item* is a hypernym of *lexical sense* and *lexical entry* as defined in Lemon-OntoLex ('Lexicon Model for Ontologies), which, in turn, correspond with *lexical unit* and *lexeme* as defined by Cruse in *Lexical Semantics*.

[20] Brewer, Review of *Historical Thesaurus of the Oxford English Dictionary*, p. 802.

[21] Kay et al., 'Unlocking the OED', p. xiv.

[22] Coleman, Review of *Historical Thesaurus of the Oxford English Dictionary*, p. 208; and Busse, 'A Celebration of Words and Ideas', p. 88., respectively.

[23] Diller, 'Emotions in the English Lexicon'.

[24] Crystal, *Words in Time and Place*.

used to denote simply 'boy, youth attendant'. Such a semantic shift is known as amelioration. The opposite of amelioration is possible as well, called pejoration. An example of this semantic shift can be found with the lexical item *knave*, meaning 'crook', which was used to indicate the more neutral meaning of 'boy, servant' in the Old English period.[25] Although such changes in meaning of lexical items may be discoverable through a historical dictionary, such as the *Oxford English Dictionary*, too, a diachronic thesaurus offers the means to track these changes through the semantic framework constituted by its topical systems.[26] Thus, gaps in meaning that such language changes came to fill in a lexicon can be traced across semantic fields. The same is true for gaps left behind by such changes. Competition between words and phrases within a semantic field is more apparent, too, by identifying and contrasting lexical items in a field that have survived or were newly coined with others in that field that disappeared from use.[27] Tracing developments in the language in such a manner is only one of the possible research uses for topical thesauri.

Another use for these lexicographic resources lies in research on metaphors.[28] By mapping out which words, or lexical items, can be used metaphorically to indicate other notions, it is possible to see which semantic fields in a language have close conceptual ties. As an example, words related to sleeping are not uncommonly used to indicate death, the everlasting sleep, as it were. Additionally, lexical items indicating temperature have a strong metaphorical link with those of emotions; a hot-headed person is one who is easily angered.[29] By mapping out metaphorical connections such as these, it is possible to gain a better understanding of the stylistic impact of metaphors and to grasp which group of words are more easily used to symbolically represent other meanings (see, for instance, Figure 0.1).

Since lexicons are culture-specific, thesauri allow one to study a culture.[30] One manner in which important cultural aspects can be studied is by noting which words exist, or existed, in a vocabulary. In Old English, for instance, the lexical item *wergild* denotes the legal value of a person's life.[31] The existence of this word, or perhaps rather the need for it, is a reflection of the fact that the penal laws of this early medieval society often required the perpetrator to recompense damages to other people by the amount of money the injured (or killed) person was deemed worth in society. The higher the rank of a person, the higher their worth.[32] Further cultural insights can be gained from thesauri by analysing what

[25] For an introduction to amelioration and pejoration, including the examples mentioned here, see Schendl, *Historical Linguistics*, p. 31.

[26] E.g., Kay and Wotherspoon, 'Wreak, Wrack, Rack, and (W)ruin'.

[27] E.g., Tejada-Caller, 'On *Shapelings* and *Childlings*'.

[28] Allan, *Metaphor and Metonymy.*

[29] For both metaphorical ties mentioned, see *Mapping Metaphor.*

[30] Kay and Roberts, *The Encyclopedia of Language and Linguistics*, s.v. 'Thesaurus'. An introduction to links between language and culture – including vocabulary expansion owing to inventions and social attitudes reflected in expressions – is provided by Kay and Allan, 'Language and Culture'. Furthermore, Chapter 8 of the dissertation reflects on *A Thesaurus of Old English*, specifically, and its relation to the culture of its early medieval speakers.

[31] Bosworth and Toller, *An Anglo-Saxon Dictionary*, s.v. 'wer-gild'.

[32] Wormald, 'Anglo-Saxon Society and its Literature', p. 11.

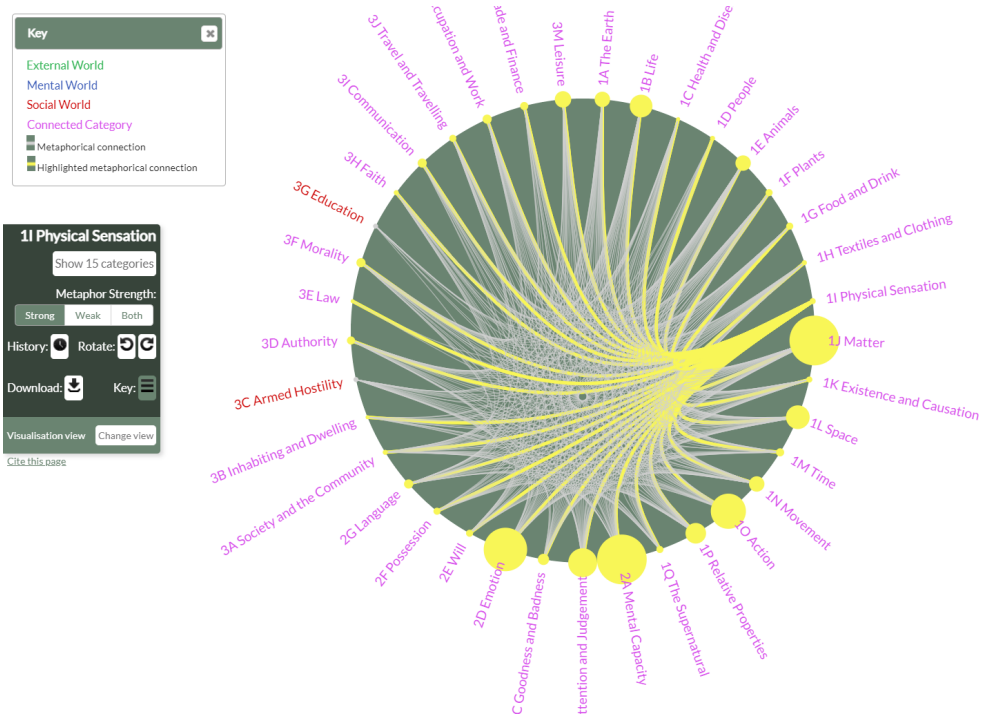


Figure 0.1.: Screenshot of the *Mapping Metaphor* website, which visualizes metaphorical connections between semantic fields.

distinctions people in a culture felt warranted reflection in their vocabulary. The group of loosely synonymous words for body in *TOE*, for instance, includes the lexical items *līc* and *līchama*. The former generally referred to dead bodies; the latter to living — a distinction early medieval English speakers felt necessary to maintain in order to express their thoughts.[33]

Moreover, the importance of concepts in a culture can be judged by their elaboration in the available vocabulary: a relatively large number of words to express related notions and the nuances between them tends to convey such significance.[34] In the surviving Old English vocabulary, for instance, the number of words to express bondage, slavery is quite sizeable, "show[ing] how this warrior society was sustained by a class of unfree [...] men and women".[35] Another striking example is that *TOE* lists thirty-six Old English words for cloak-like garments. The great variety in words available to the Anglo-Saxons to describe these garments "suggests that cloaks were a common garment, worn by different social classes".[36]

Next to cultural research, thesauri are also of great value for literary-critical research. By annotating literary works with definitions in a lexicographic work, it is possible to indicate, per word, which specific meaning an author must have intended to convey. Cases of ambiguity can be made explicit by referencing

[33] Bosworth and Toller, *An Anglo-Saxon Dictionary*, s.v. 'líc' and 'líc-hama'.
[34] Wierzbicka, *Understanding Cultures through their Key Words*, pp. 10–11.
[35] Momma, Review of *A Thesaurus of Old English*, p. 80.
[36] *Learning with the Online* Thesaurus of Old English, section 'Unit 5 Clothing'.

multiple such definitions. Doing so may aid scholars in discussions on what the intentions and implications are of a literary work. Applying definitions from a thesaurus rather than from a dictionary has the advantage that scholars can, through the topical system of the thesaurus, obtain further insights into the diction and stylistic choices made by authors. Which lexical items were available to Shakespeare? Which ones seem to have been restricted to poetry (possibly because they were considered archaic), and do such restrictions appear to have had an impact on the word choices Shakespeare made for his plays? Which words did the poet and playwright prefer over others that expressed the same notion? In fact, answering questions on stylistic preference may help in establishing the author of a text.[37] By applying statistical analyses, using thesauri to deduce which alternatives would have been available for a particular word, the deduced information on preferences could lead to establishing authorial fingerprints, as it were, which may assist in identifying which other literary works likely belong to the same author.[38]

In short, not only have thesauri proven themselves to be major assets in everyday use, they have been shown to be of great worth in various fields of research — research not limited to linguistics, but extending to cultural, anthropological, and literary-critical research.

0.3. Digital research opportunities

While most thesauri were fashioned to categorise contemporaneous vocabularies,[39] very few have been fashioned with a historical language as their subject.[40] The lack of such thesauri – in spite of their abundant and diverse uses – is straightforward to explain. Developing and publishing the first edition of *TOE*, for instance, has taken a team of researchers over fifteen years.[41] *HTE*, covering a vastly larger time period, has taken a team of two hundred thirty people well over forty years to develop.[42] These staggering numbers are certain to discourage most scholars from attempts at manually fashioning comprehensive thesauri of other historical languages. Future research would therefore benefit greatly from means

[37] An example of lexical evidence towards establishing the author of an anonymous text can be found in Bremmer, 'Old English "Cross" Words'. In surveying Old English words denoting Christ's cross, Bremmer concludes that the word *rōdehengen* must have been Ælfric of Eynsham's "own coinage" (p. 220) and strengthens the claim that the anonymous, liturgical text in which this word is found, too, is by the hand of this medieval abbot. A further exploration and analysis of Ælfric's lexical preferences, utilizing a thesaurus, can be found in Van Baalen, 'Identifying, Categorising and Exploring "Ælfrician" Vocabulary'.

[38] This approach is not unlike research done in the Lexomics programme (see 'Lexomics'), although there no advantage is taken yet of thesauri for their knowledge on alternative phrasings that may have been available to a specific author.

[39] E.g., McArthur, *Longman Lexicon of Contemporary English*; Wilkinson, *Thesaurus of Traditional English Metaphors*.

[40] See Chapter 1.

[41] Roberts, 'Towards an Old English Thesaurus'.

[42] These statistics can be found in the section 'Stats and Figures' of *HTE*. Unsurprising for such vast projects, the duration can partly be attributed to funding issues (Kay et al., 'Unlocking the OED', p. xv).

to develop these resources in an automated manner and, arguably more important, to improve the dissemination form of existing thesauri in order to elaborate on and reuse the knowledge within. Such an improved dissemination may facilitate novel analyses and reuse of topical systems for other, related languages.[43] In fact, improving the form and method of dissemination of these thesauri should offer opportunities for novel research.

Reviews of *TOE* foreground a number of opportunities for improving the dissemination of historical language thesauri.[44] A prominent shortcoming of current editions is the inability to query and reuse the information contained in the thesauri in a way other than its editors had foreseen. The digital versions of *TOE* and *HTE* employ database-technology based on tabular storage of their data.[45] This data is subsequently shared, or rather visualised, in a searchable and browsable manner. However, this way of sharing proves to be rather limited. Users are unable to query or visualize these datasets in another manner than those provided for by the existing user interfaces of *TOE* and *HTE*.

An example of a welcome query to be run against thesaurus content (a query that *TOE* and *HTE* currently do not allow) is one that provides statistics on the sizes of its semantic fields. As Anna Wierzbicka notes, important concepts in a society often witness cultural elaboration; that is, the availability of a relatively large number of words to express related notions and the nuances between them.[46] Supplying users with such key statistics is straightforward within a digital environment, without them having to count the lexical items manually. For such functionality on statistics and other aspects to be added by parties other than the publisher, thesauri content should be reusable, in a storage format that does not favour one perspective (or query) over another.

Another area in which the dissemination of thesauri can be improved is in the availability of tagging information. In *TOE*, for instance, indications of date and dialect are notably absent.[47] As it stands, all items are treated as belonging to "a single geographically and temporally indistinguishable mass".[48] Moreover, expanding thesauri with such information per lexical item will allow scholars to create subthesauri: thesauri filtered to display only those lexical items that meet a set of given criteria. For a diachronic thesaurus this could be a selection of a smaller time frame within the covered period. The editors of *HTE*, for instance, point out that their thesaurus could be used to determine which lexical items will have been available to Shakespeare.[49] However, the website does not allow the creation of a subthesaurus containing only those lexical items that were

[43] See the expand method as utilised by, for instance, Miháltz et al., 'Methods and Results of the Hungarian WordNet Project'; and by Fernández-Montraveta et al., 'The Spanish Version of WordNet 3.0'.

[44] The following reviews of *TOE* have been consulted: Bremmer, Cavill, Conner, Dance, Görlach, Momma, Van Gelderen.

[45] Both *TOE* and *HTE* employ MySQL database technology, as stated in the section 'Creation of the *Thesaurus*' of *TOE*.

[46] Wierzbicka, *Understanding Cultures through their Key Words*, pp. 10–11.

[47] Bremmer, 'Treasure Digging', p. 111; Görlach, Review of *A Thesaurus of Old English*, p. 399; Dance, Review of *A Thesaurus of Old English*, p. 313.

[48] Dance, Review of *A Thesaurus of Old English*, p. 313.

[49] Kay et al., 'Unlocking the OED', p. xiv.

available during the life of Shakespeare — or any other subthesaurus for that matter. Any filtering is left as exercise to the user. *TOE*, too, includes valuable tagging information, stating whether its lexical items are found only in poetry or only in glosses. Although a previous, digital edition of *TOE* allowed listing all lexical items tagged with a specific flag, the current edition no longer sports such helpful filtering capabilities.[50] The current digital versions of these thesauri do not facilitate researchers in extending existing content with further salient information and, based on available tags, in scrutinizing only those items deemed of interest to them. In essence, the desire for extra tagging information – just like the need to reuse thesauri data in other projects – boils down to the need for extendibility: allowing further information to be contributed, making it possible to form new queries over the combined information, and ensuring visualisations will convey the new insights acquired by the extensions.

In short, current forms of historical language thesauri limit their utilization. An attempt to resolve some of these issues, taking into account recent developments in information technology and improving on existing specifications for sharing linguistic information on the Semantic Web, should open up these lexicographic resources further for novel research.

0.4. Research objective

Historical language thesauri offer a number of uses beyond discovering available alternative phrasings, but, as explained in the previous section, opportunities exist to further their utilization for research. The main objective of my dissertation is to explore these opportunities and improve the use of these valuable lexicographic resources across various disciplines in academia. The result should enable a wider use of these thesauri for cultural, linguistic, anthropological, and literary-critical research.

0.4.1. Research questions

The main question of my research is formulated as follows:

> How can Web-based dissemination of thesauri of historical languages, and thesauri in general, be improved so as to answer to the research needs of scholars in various disciplines?

This overarching research question is covered by three sub questions on historical language thesauri:

1. What are the main components found in these thesauri?
2. What are the main features, or functionality, of these thesauri that are desired for research?

[50] For a description of the *TOE* website previously available and its filtering on tagging information, see Stolk, 'Welcoming the *Thesaurus of Old English Statistics*', pp. 11–14.

3. What digital form should these thesauri be published in on the Web – and what modifications to current specifications ought to be implemented for this purpose – to support a wider use of historical language thesauri in academia?

The questions above are addressed in Part I and II of the dissertation. The answers and hypotheses yielded are subsequently adopted and evaluated, in Part III, through their application to the historical language thesaurus *TOE*.

0.5. Dissertation outline

This section provides an overview of the various chapters which, together, aim to answer the research questions. Besides the introduction and conclusion, the dissertation contains nine chapters – i.e., three peer-reviewed papers, two peer-reviewed articles, and four original chapters – that are spread across three Parts.[51] These chapters are discussed in terms of their position within the overall dissertation, their relation to the research questions, and their approach and most notable findings.

Part I. Historical Language Thesauri and their Characteristics

Answering the main research question demands an understanding of what historical language thesauri are and what needs researchers have in accessing them. Part I of the dissertation consists of two original chapters, which provide insight into these matters through an overview of the characteristics of historical language thesauri.

Chapter 1 addresses research sub question 1: "What are the main components found in historical language thesauri?" This chapter draws from two types of sources in order to provide an overview of the information found in thesauri: (1) existing historical language thesauri of Scots and English and (2) publications and handbooks on both thesauri and lexicography in general. In the analysis and resulting overview, the focus lies on the knowledge contained within thesauri rather than at how that content is presented. Knowledge on the former can be used to produce multiple different presentations of the same thesaurus content, whilst the latter would mostly serve a single form of presentation and varies between different publications of a thesaurus (e.g., print editions and online editions). The three main parts of thesauri distinguished are: (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.

Chapter 2 addresses research sub question 2: "What are the main features, or functionality, of historical language thesauri that are desired for research?" This

[51] The dissertation contains references to code (of both software and data transformations) that has been developed and published as part of this research. An overview thereof is provided in the 'List of source code' in the back matter of the dissertation.

chapter consults, in addition to the sources for Chapter 1, academic reviews of these thesauri and notable research employing them. These sources help establish which aspects of the existing historical language thesauri are deemed an asset, and which were found wanting or absent. Further input was gathered through a series of workshops surrounding a single thesaurus (*TOE*) in order to obtain further research needs. The resulting information on desired functionality – i.e., navigation, resource views, extension, analyses, and data management – has been translated to a set of requirements for the digital form proposed for thesaurus publications on the Web and for the web application developed for interacting with these lexicographic resources.

Part II. Historical Language Thesauri and a Digital Form on the Semantic Web

Acknowledging the determined characteristics, both current and desired, of historical language thesauri, Part II explores a suitable digital form for their publication on the Web. These explorations are covered by one original chapter and two peer-reviewed publications.

Chapter 3 reflects on existing Web-based publications of historical language thesauri, discusses best practices for publishing data on the Web, and proposes a digital form for these thesauri that may be more appropriate for the research needs in mind than those of existing publications. The digital form proposed, which adopts Linked Data paradigms, facilitates data being FAIR (findable, accessible, interoperable, reusable). Through an analysis of each typical information component found in thesauri, a combination of suitable Linked Data data vocabularies is constructed and advocated in which to express thesauri. The combination and use of these vocabularies has, in a wider community towards standardization of the representation of linguistic and lexicographic resources on the Web, recently been termed Linguistic Linked Data.

The next two chapters are peer-reviewed papers published as part of conference proceedings, both of which offer findings on one of the core data vocabularies of Linguistic Linked Data, Lemon-OntoLex, and its applicability to thesauri. *Chapter 4*, entitled 'OntoLex and Onomasiological Ordering: Supporting Topical Thesauri', was published as part of the proceedings of the conference 'Language, Data, and Knowledge 2017', Galway. This paper argues that Lemon-OntoLex is, on its own, insufficient for representing a large proportion of existing thesauri. After demonstrating the current expressivity and mentioned shortcoming of Lemon-OntoLex through two case studies (i.e., *The Historical Thesaurus of the Oxford English Dictionary* and *The Scots Thesaurus*), the paper proposes the addition of a relation to the OntoLex vocabulary: `ontolex:isSenseIn`. The findings and proposal of this paper was further discussed within the OntoLex community and resulted in the creation of a complementary data vocabulary, *lemon-tree*, which is treated in the second paper included in the dissertation.[52]

Chapter 5, a peer-reviewed paper entitled '*lemon-tree*: Representing Topical

[52] Based on feedback from the OntoLex community, the proposed `isSenseIn` was included in the *lemon-tree* model as `isSenseInConcept`.

Thesauri on the Semantic Web', was published as part of the conference proceedings of 'Language, Data, and Knowledge 2019', Leipzig. This paper analyses fundamental needs for representing thesauri as Linguistic Linked Data and offers a solution for capturing two important aspects not covered by Lemon-OntoLex: (1) levels that can be distinguished in the topical system of thesauri and (2) a looser form of categorization than lexicalization. The necessary terminology for these aspects has been made available in an information model called *lemon-tree*, facilitating publications of thesauri as Linguistic Linked Data.

Part III. Disseminating and Evaluating *A Thesaurus of Old English* as Linguistic Linked Data

The FAIR digital form of thesauri on the Web, proposed in Part II, is evaluated in Part III. Here, the case study of a single historical language thesaurus, *TOE*, is discussed in-depth: its transformation to the digital form, dissemination through the novel web application Evoke, and utilization in research across various disciplines in academia. Thus, this part assesses to what extent Linguistic Linked Data, and the manner in which it is disseminated, allows for a wider use in academia than existing paper and digital editions of historical language thesauri. Part III contains three peer-reviewed publications and one original chapter.

Chapter 6 is a paper, entitled '*A Thesaurus of Old English* as Linguistic Linked Data: Using OntoLex, SKOS and *lemon-tree* for Bringing Topical Thesauri to the Semantic Web', presented at 'eLex 2019', Sintra. This paper discusses the process of porting *TOE* to the Web-based form and provides recommendations for representing topical thesauri on the Web, whilst granting insights into aspects that may be encountered in porting similar lexicographic resources in the future. In order to support future work on other thesauri, the automated process created for porting this thesaurus has been made available via open access, too.[53]

Chapter 7 is the article 'Evoke: Exploring and Extending *A Thesaurus of Old English* using a Linked Data Approach'. Published in the international peer-reviewed journal *Amsterdamer Beiträge zur älteren Germanistik* in 2021, this article details the web application Evoke. This web application, developed as part of this research, offers functionality for navigating, viewing, extending, and analysing thesaurus content that is represented as Linguistic Linked Data. Figure 0.2 illustrates the ability to navigate the topical system of a historical language thesaurus. So-called breadcrumbs indicate the current location in the topical system: starting from "Power, might" down to, currently in view, "Freedom, being free". The large pane in the user interface here indicates which subordinate categories, such as "A free man" and "A free woman", are available to the user for navigating further down the semantic hierarchy that the topical system constitutes. The functionality of Evoke, including the means to navigate, is founded on the research needs that surfaced in Part I of the dissertation and in various workshops referred to earlier. As the article demonstrates, Evoke proves to be a powerful research tool that facilitates its users to perform novel cultural linguistic analyses over multiple sources.

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

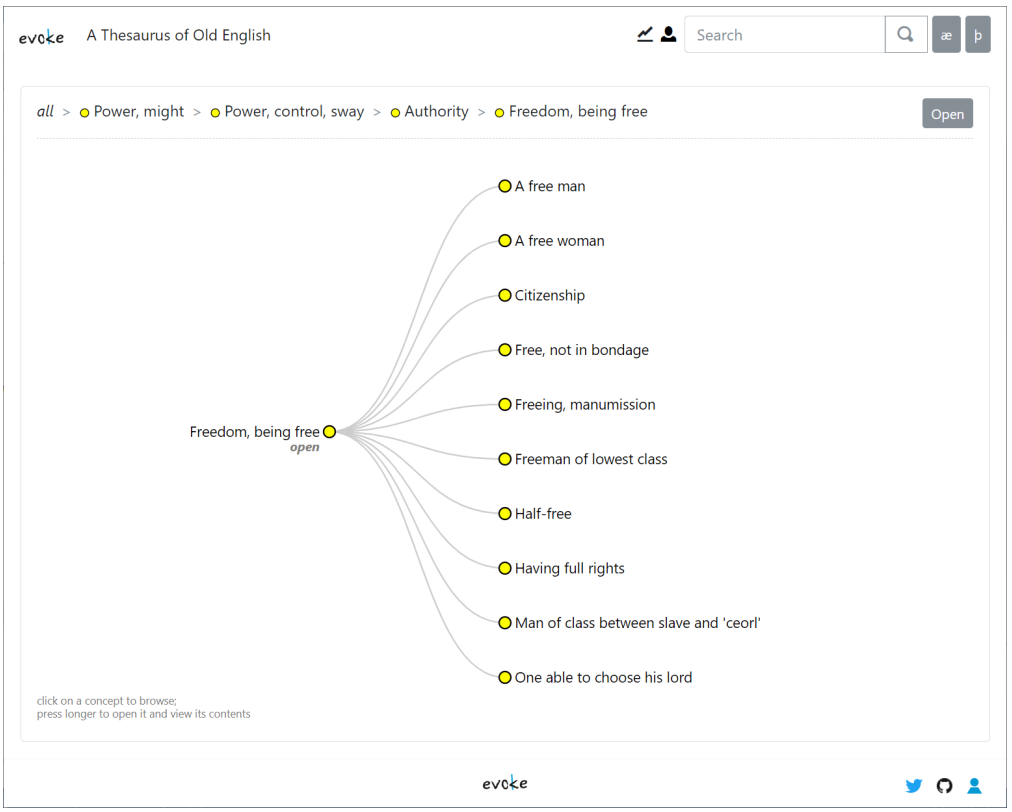


Figure 0.2.: Navigating the topical system of *TOE* in the web application Evoke.

Chapter 8, following the chapter on Evoke, assesses the usefulness of the dissemination of *TOE* as Linguistic Linked Data through the discussion of a number of research case studies. A collaborative project, titled 'Exploring Early Medieval English Eloquence' (EEMEE), was established for this purpose. The project has brought together scholars from universities and lexicographic institutions from across Europe to explore – and elaborate on – the contents of *TOE* using the web application Evoke. Their case studies, the majority of which have been published as part of a special issue of the international, peer-reviewed journal *Amsterdamer Beiträge zur älteren Germanistik*, are summarized in this chapter, along with reflections on the benefits and disadvantages of the new digital form and dissemination of the thesaurus concerned.

Chapter 9 is an article on one of the research case studies mentioned in Chapter 8, co-written by Rita van de Poel and myself, titled 'A Case of Kinship: Onomasiological Explorations of KINSHIP in Old Frisian and Old English'. Explorations of Old Frisian and Old English lexis in the semantic field of KINSHIP are realized by connecting Old Frisian lexis to the overarching structure of *TOE*. The resulting dataset for KINSHIP in Old Frisian thus shares its semantic framework with that for Old English lexis. As the article demonstrates, this approach facilitates comparative analyses of the two historical languages, on an onomasiological level, which leads to new insights into linguistic and cultural aspects of these two languages and their language communities.

Lastly, the dissertation concludes by providing answers to the research questions, reflecting on the extent with which the proposed dissemination form of historical language thesauri overcomes the listed impediments for research, and discussing the potential future impact of the research presented here.

0.6. Related work

Digital approaches for investigating various facets surrounding language, including the structure of vocabularies and diachronic perspectives on language development, are by no means new.[54] Such matters have often been explored, in various branches of computational linguistics, with digital corpora and analytical tools.[55] Onomasiological approaches are included amongst them, most notably through the use of thesauri and smaller semantic field studies.[56] Since the last century, research programmes have sought to further harness the potential of digital thesauri, whether digitized or born-digital. Automated uses, taking advantage of their digital form, include natural language processing and automated translations.[57] The next paragraphs will position my dissertation within a number of important developments within the field of Digital Humanities that deal with the study of lexicography and onomasiology.

The research in my dissertation shares its aim of increasing the availability of research infrastructure for language resources with recent European initiatives. Notable infrastructures for the Humanities are CLARIN (Common Language Resources and Technology Infrastructure)[58] and DARIAH (Digital Research Infrastructure for the Arts and Humanities)[59]. These infrastructures allow researchers to work with, amongst others, textual corpora and natural language processing. Topical thesauri, the focus in this study, are not included in these two infrastructures. Indexing thesauri, however, are supported and are published as Linked Data.[60] This format, which is also adopted in my dissertation, has been refined and standardized for linguistic resources in the Linguistic Linked Data community.[61] In fact, CLARIN and the Linguistic Linked Data community seek to strengthen collaborations and have discussed common goals.[62]

An important context for the research presented in the dissertation, as alluded to in the previous paragraph, is the efforts surrounding Linguistic Linked Data. As Declerck et al. have observed, these efforts play "an increasing role in eLexicography".[63] The English WordNet, for instance, has recently

[54] This section draws on the overview of related work as published in Stolk, 'Evoke'.
[55] Sula and Hill, 'The Early History of Digital Humanities'.
[56] *HTE* lists many such studies in its 'Bibliography' section.
[57] See, for instance, the use of machine-learning based on BERT in Kohli, 'Transfer Learning and Augmentation for Word Sense Disambiguation'.
[58] De Jong et al., 'Interoperability in an Infrastructure Enabling Multidisciplinary Research'.
[59] Maryl et al., 'Future of Scholarly Communication'
[60] 'Vocabs Services', *DARIAH*.
[61] Cimiano et al., *Linguistic Linked Data*.
[62] Stokman, 'A Recap on the CLARIN Café on Linguistic Linked Data', *CLARIN*.
[63] Declerck et al., 'Recent Developments for the Linguistic Linked Open Data Infrastructure', p. 5664.

been ported to this model.[64] Moreover, several recent initiatives aim at building and maintaining Linguistic Linked Data resources, including the H2020 projects ELEXIS (2018–22), Prêt-à-LLOD (2019–22) and the COST Action NexusLinguarum (2019–23).[65] Tooling in these initiatives that work with Linguistic Linked Data focus on creation, discovery, transformation, and linking.[66] Examples of such tools include LingHub, which offers discovery of language resources by searching through their metadata,[67] and NAISC, used for aligning two RDF datasets. Unfortunately, most applications currently available for working with Linguistic Linked Data "come with a considerable entry barrier and they address the advanced user of RDF technologies rather than a typical linguist".[68] The web application Evoke, developed as part of my dissertation, is amongst the first range of applications that aims to provide a user-friendly interface for such resources and to open them up to a wider audience. Other notable applications that provide user interfaces for Linguistic Linked Data resources are VocBench 3 and LexO.[69] Both of these web-based platforms allow users to edit and view Linguistic Linked Data in a user-friendly manner. However, unlike Evoke, they lack functionality to perform onomasiological analyses: their main aim is to manage and publish content collaboratively.

Lastly, a number of recent research programmes have increased efforts that expand the use of thesauri to other domains. The onomasiological lens that *HTE* provides, for instance, has been utilized for mapping metaphors throughout the history of the English language and for semantically annotating entire textual corpora for topical analyses.[70] Similarly, my dissertation seeks to contribute novel methods to Digital Humanities research for engaging with thesauri. By offering statistical analyses utilizing the semantic hierarchy of these lexicographic resources, and by allowing researchers to link additional information to thesaurus content, the web application Evoke grants new, meaningful insights into a language and the use of its vocabulary in cultural expressions (e.g., individual texts or entire oeuvres). As Chapter 8 demonstrates, the functionality available offers results that provide additional knowledge, but may also raise new questions that warrant a closer inspection of the cultural context (e.g., textual, historical, socio-economic). The research presented here, therefore, is firmly rooted in Digital Humanities, and provides the means to explore Humanities-based questions through digital resources that complement knowledge and expertise of scholars.

[64] McCrae et al., 'English WordNet 2020'.

[65] ELEXIS: https://cordis.europa.eu/project/id/731015, 2018–2022; Prêt-à-LLOD: https://cordis.europa.eu/project/id/825182, 2019–2022; NexusLinguarum: https://www.cost.eu/actions/CA18209, 2019–2023.

[66] Declerck et al., 'Recent Developments for the Linguistic Linked Open Data Infrastructure'.

[67] McCrae and Cimiano, 'Linghub'.

[68] Chiarcos et al., 'On the Linguistic Linked Open Data Infrastructure', p. 13.

[69] Stellato et al., 'VocBench 3'; Bellandi and Giovannetti, 'Involving Lexicographers in the LLOD Cloud with LexO'.

[70] *Mapping English Metaphor through Time*; Piao et al., 'A Time-sensitive Historical Thesaurus-based Semantic Tagger for Deep Semantic Annotation'.

References

Adamska-Salaciak, A., Review of *Historical Thesaurus of the Oxford English Dictionary* by C. Kay et al., *International Journal of Lexicography* 23.2 (2010), 227–33.

Allan, K., *Metaphor and Metonymy: A Diachronic Approach*, Publications of the Philological Society 42 (Malden, 2008).

Bellandi, A. and E. Giovannetti, 'Involving Lexicographers in the LLOD Cloud with LexO, an Easy-to-Use Editor of Lemon Lexical Resources', Proceedings of the 7th Workshop on Linked Data in Linguistics (2020), 70–4. https://aclanthology.org/2020.ldl-1.10.pdf.

Bosworth, J. and T. N. Toller, *An Anglo-Saxon Dictionary Based on the Manuscript Collections of the Late Joseph Bosworth* (London, 1898), *Supplement by T. N. Toller* (Oxford, 1921), *with Enlarged Addenda and Corrigenda by A. Campbell* (Oxford, 1972).

Bremmer Jr, R. H., 'Treasure Digging in the Old English Lexicon', Review of *A Thesaurus of Old English* by J. Roberts et al., *NOWELE* 40 (2002), 109–14.

Bremmer Jr, R. H., 'Old English "Cross" Words', *Cross and Cruciform in the Anglo-Saxon World: Studies to Honor the Memory of Timothy Reuter*, Medieval European Studies XI (Morgantown, 2010), 204-32.

Brewer, C., Review of *Historical Thesaurus of the Oxford English Dictionary* by C. Kay et al., *Review of English Studies* 61.252 (2010), 801–5.

Busse, B., 'A Celebration of Words and Ideas: The Stylistic Potential of the *Historical Thesaurus of the Oxford English Dictionary*', *Language and Literature* 21.1 (2012), 84–92.

Cavill, P., 'Names and Things in Anglo-Saxon and Early Norman England', Review of *A Thesaurus of Old English* by J. Roberts et al. and *Words, Names and History: Selected Writings of Cecily Clark* by P. Jackson, *Nottingham Medieval Studies* 41 (1997), 186–91.

Celen, K. M. and Ş. Yalçin, 'The Effects of Vocabulary Resource Use on Lexical Richness in L2 Writing', *Milli Eğitim Dergisi* 50.230 (2021), 1039-58. doi: 10.37669/milliegitim.687806.

Chiarcos, C. et al., 'On the Linguistic Linked Open Data Infrastructure', Proceedings of the 1st International Workshop on Language Technology Platforms (2020), 8–15. https://www.aclweb.org/anthology/2020.iwltp-1.2.

Cimiano, P. et al., *Linguistic Linked Data: Representation, Generation and Applications* (2020). doi: 10.1007/978-3-030-30225-2.

Coleman, J., *Love, Sex and Marriage: A Historical Thesaurus*, Costerus New Series 118 (Amsterdam, 1999).

Coleman, J., Review of *Historical Thesaurus of the Oxford English Dictionary* by C. Kay et al., *Word Structure* 6.2 (2013), 201–13.

Collins English Thesaurus, eds. J. Crozier et al. (Glasgow, 2011).

Conner, P. W., Review of *A Thesaurus of Old English* by J. Roberts et al., *Speculum* 73.3 (1998), 887–9.

Crystal, D., *Words in Time and Place: Exploring Language through the Historical Thesaurus of the* Oxford English Dictionary (Oxford, 2014).

Dance, R., Review of *A Thesaurus of Old English* by J. Roberts et al., *Medium Ævum* 66.2 (1997), 312–3.

Declerck, T. et al., 'Recent Developments for the Linguistic Linked Open Data Infrastructure', Proceedings of the 12th Language Resources and Evaluation Conference (2020), 5660–7. https://www.aclweb.org/anthology/2020.lrec-1.695.

De Jong, F. et al., 'Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN', Proceedings of the 12th International Conference on Language Resources and Evaluation (2020). https://www.aclweb.org/anthology/2020.lrec-1.417/.

Diller, H., 'Emotions in the English Lexicon: A Historical Study of a Lexical Field', *English Historical Linguistics* 1992: Papers from the 7th International Conference on English Historical Linguistics, eds. F. Fernandez et al., Current Issues in Linguistic Theory 113 (Amsterdam, 1994), 219–34.

Encyclopedia of Language & Linguistics, 2nd edn, ed. K. Brown, 14 vols. (Oxford, 2006).

'EuroVoc', *EUR-Lex*. https://eur-lex.europa.eu/browse/eurovoc.html. Accessed: 15 March 2022.

Faria, F., 'Georges Cuvier et le premier paradigme de la paléontologie', *Revue de Paléobiologie* 32.2 (2013), 297–302.

Fernández-Montraveta, A., G. Vázquez, and C. Fellbaum, 'The Spanish Version of WordNet 3.0', Text resources and Lexical Knowledge', *Text, Translation, Computational Processing* (Berlin, 2008), 175–82.

Gánem-Gutiérrez, G. A. and A. Gilmore, 'A Mixed Methods Case Study on the Use and Impact of Web-based Lexicographic Tools on L2 Writing', *Computer Assisted Language Learning* 2021. doi: 10.1080/09588221.2021.1987273.

Görlach, M., Review of *A Thesaurus of Old English* by J. Roberts et al., *Anglia* 116.3 (1998), 398–401.

HTE = The Historical Thesaurus of English, eds. C. Kay et al. https://ht.ac.uk/. Accessed: 15 March 2022.

Hüllen, W., *A History of Roget's* Thesaurus*: Origins, Development, and Design* (Oxford, 2005).

Kay, C. and I. Wotherspoon, 'Wreak, Wrack, Rack, and (W)ruin: The History of Some Confused Spellings', in Sounds, Words, Texts and Change: Selected Papers from 11 ICEHL, Santiago de Compostela, 7-11 September 2000, 129-43 (Amsterdam, 2002).

Kay, C. and K. Allan, 'Language and Culture', in *English Historical Semantics*, Edinburgh Textbooks on the English Language - Advanced (Edinburgh, 2015).

Kay, C. and M. Alexander, 'Diachronic and Synchronic Thesauruses'. *The Oxford Handbook of Lexicography* (Oxford, 2016), 367–80.

Kay, C. et al., *Historical Thesaurus of the Oxford English Dictionary* (Oxford,

2009).

Kay, C. et al., 'Unlocking the *OED*: The Story of the *Historical Thesaurus of the OED*', *Historical Thesaurus of the Oxford English Dictionary*, eds. C. Kay et al. (Oxford, 2009), xiii–xx.

Kohli, H., 'Transfer Learning and Augmentation for Word Sense Disambiguation', Proceedings of the 43rd European Conference on Information Retrieval (2021), 303–11. doi: 10.1007/978-3-030-72240-1_29.

Learning with the Online Thesaurus of Old English, eds. C. Hough and C. Kay. https://oldenglishteaching.arts.gla.ac.uk. Accessed: 15 March 2022.

'Lexomics', *Wheaton College*. http://wheatoncollege.edu/lexomics/. Accessed: 22 February 2016.

Mapping English Metaphor through Time, eds. W. Anderson, E. Bramwell, and C. Hough, (Oxford, 2016).

Mapping Metaphor. http://mappingmetaphor.arts.gla.ac.uk/. Accessed: 15 March 2022.

Maryl, M. et al., 'Future of Scholarly Communication: Forging an Inclusive and Innovative Research Infrastructure for Scholarly Communication in the Social Sciences and Humanities', Research Report, OPERAS (2021).

McArthur, T., *Longman Lexicon of Contemporary English* (Harlow, 1981).

McCrae, J. et al., 'English WordNet 2020', Proceedings of the Multimodal Wordnets Workshop at LREC 2020 (2020), 14–19.

McCrae, J. and P. Cimiano, 'Linghub: A Linked Data Based Portal Supporting the Discovery of Language Resources', Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems (Vienna: 2015), 88–91. http://dblp.uni-trier.de/db/conf/i-semantics/semantics2015p.html#McCraeC15.

'Medical Subject Headings RDF', *U.S. National Library of Medicine*. https://id.nlm.nih.gov/mesh/. Accessed: 15 March 2022.

Mihăltz, M. et al., 'Methods and Results of the Hungarian WordNet Project', Proceedings of the 4th Global WordNet Conference (Szeged, Hungary, 2008), 311–20.

Momma, H., Review of *A Thesaurus of Old English* by J. Roberts et al., *Notes and Queries* 50.1 (2003), 79–80.

NAISC. https://github.com/insight-centre/naisc. Accessed: 15 March 2022.

'NASA Thesaurus', *Data.gov*. https://catalog.data.gov/dataset/nasa-thesaurus. Last updated: 12 November 2020.

Oxford English Dictionary. https://www.oed.com. Accessed: 15 March 2022.

Piao, S. et al., 'A Time-sensitive Historical Thesaurus-based Semantic Tagger for Deep Semantic Annotation', *Computer Speech and Language* 46 (2017), 113–35.

Roberts, J., 'Towards an Old English Thesaurus', *Poetica* 9 (1978), 56–72.

Roberts, J., '*A Thesaurus of Old English*: The Pilot Study for the Glasgow *Historical Thesaurus*', *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), 297–317. doi: 10.1163/18756719-12340234.

Roget's = Roget, P. M., *Thesaurus of English Words and Phrases, Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in*

Literary Composition (London, 1852).
Schendl, H., *Historical Linguistics, Oxford Introductions to Language Study* (Oxford, 2001).
'Stellato, A. et al., 'VocBench 3: A Collaborative Semantic Web Editor for Ontologies, Thesauri and Lexicons', *Semantic Web* 11.5 (2020), 855–81.
Stolk, S., 'Welcoming the *Thesaurus of Old English Statistics*: The *Thesaurus of Old English* and the Vocabulary of Greetings', M.A. dissertation, Leiden University (2013).
Stolk, S., 'OntoLex and Onomasiological Ordering: Supporting Topical Thesauri', Proceedings of the Linked Data Knowledge 2017 Workshops, NUI Galway, 18 June 2017, pp. 60–7. http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_3.pdf.
Stolk, S., '*lemon-tree*: Representing Topical Thesauri on the Semantic Web', Proceedings of Linked Data Knowledge 2019, Leipzig, 20–23 May 2019. doi: 10.4230/OASIcs.LDK.2019.16.
Stolk, S., '*A Thesaurus of Old English* as Linguistic Linked Data', Proceedings of eLex2019, Sintra, 1–3 October 2019. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_13.pdf.
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), 318–58. doi: 10.1163/18756719-12340235.
Stokman, L., 'A Recap on the CLARIN Café on Linguistic Linked Data', *CLARIN*. https://www.clarin.eu/news/recap-clarin-cafe-linguistic-linked-data. Last updated: 17 May 2021.
Sula, C. A. and H. V. Hill, 'The Early History of Digital Humanities: An Analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004)', *Digital Scholarship in the Humanities* 34 (2019), i190–206. doi: 10.1093/llc/fqz072.
Tejada-Caller, P., 'On *Shapelings* and *Childlings*: A Linguistic Approach to the Emergence of New Cultural Borders Between the Unborn and the Newborn Child in EME (1500-1700)', *Journal of English Studies* 18 (2020), 227-51. doi: 10.18172/jes.4502.
Thesaurus Linguae Latinae, ed. M. Hillen. https://thesaurus.badw.de/en/. Accessed: 15 March 2022.
The Encyclopedia of Language and Linguistics, ed. R. E. Asher, 10 vols. (Oxford, 1994).
The Scots Thesaurus, eds. I. Macleod et al. (Aberdeen, 1990).
TOE = A Thesaurus of Old English, eds. J. Roberts et al. (Glasgow, 2015). http://oldenglishthesaurus.arts.gla.ac.uk. Accessed: 15 March 2022.
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.
Van Gelderen, E., Review of *A Thesaurus of Old English* by J. Roberts et al., *Studies in Language* 27.1 (2003), 200–3.
'Vocabs Services', *DARIAH*. https://www.dariah.eu/tools-services/tools-and-services/tools/vocabs-services/.
Wierzbicka, A., *Understanding Cultures through their Key Words: English, Russian, Polish, German, and Japanese*, Oxford Studies in Anthropological Linguistics 8, ed. W. Bright (Oxford, 1997).
Wilkinson, P. R., *Thesaurus of Traditional English Metaphors*, 2nd edn (London, 2002).
Wormald, P., 'Anglo-Saxon Society and its Literature', *The Cambridge Companion to Old English Literature*, eds. M. Godden and M. Lapidge (Cambridge, 1991), 1–22.

PART I

Historical Language Thesauri and Their Characteristics

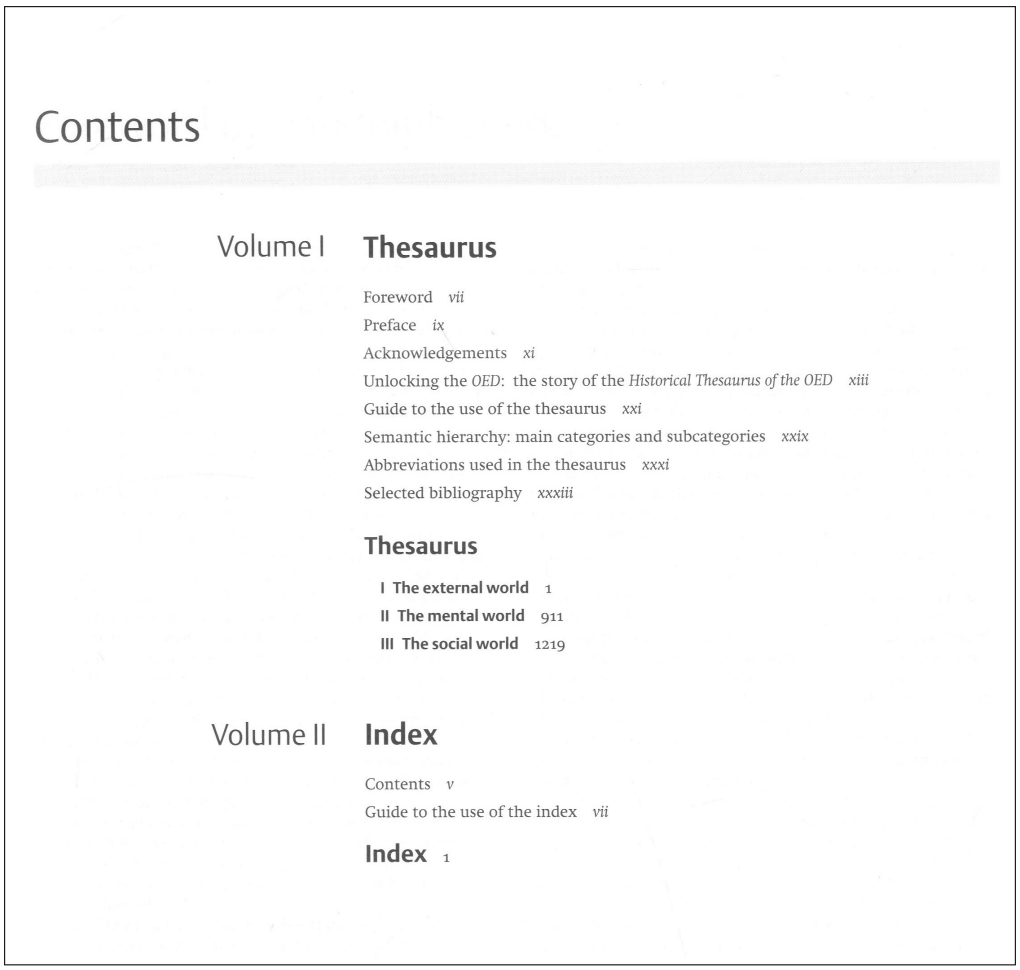
Chapter 1

1. The content of historical language thesauri

A thorough understanding of the contents of historical language thesauri is essential in determining an appropriate dissemination form for these resources. To this end, the current chapter discusses their main components and discerns generic characteristics of their information. The table of contents of the print edition of the *Historical Thesaurus of English* (*HTE*), shown in Figure 1.1, lends itself to defining what, for the purposes of this chapter, is considered the content of a thesaurus. The table applies the term *thesaurus* to the following notions: (1) the entire, two-volume publication of the thesaurus, (2) the first volume of the publication, of which the largest part is called by the same name, namely (3) the thesaurus proper. The current chapter will discuss the contents of thesauri in this narrowest of meanings, that of the thesaurus proper. In this discussion, the aim is to understand what knowledge these thesauri contain. Through an analysis of their presentations, which may vary between different publications of the same thesaurus (e.g., print editions and online editions), the chapter distils the meaning editors have attempted to convey to users.

In order to provide an overview of the information found in historical language thesauri, the chapter draws from two main types of sources. The first, and most important kind of source, is the historical language thesauri themselves. This chapter analyses thesauri of Scots and English, specifically. Both their content proper and their introductions provide valuable insights into the main components found within historical language thesauri. The second group of sources is formed by publications and handbooks on both thesauri and lexicography in general. These secondary sources impart a broader understanding of the scope and context of historical language thesauri on the one hand and, on the other, delve more deeply into the semantics of some of the elements found in these thesauri.

The remainder of the chapter is laid out as follows. Section 1.1 lists available editions of historical language thesauri of Scots and English. Subsequently, section 1.2 identifies the main components found in these historical language thesauri. The traditional lexicographic perspective, which is based on the presentation of thesauri, is argued to be less suitable for electronic editions than for paper ones and is therefore supplanted by an alternative division into components — one based on the knowledge that presentations convey. The three components identified thus, treated separately in sections 1.3–1.5, are: (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 (3) relations of synonymy, indicated through groupings of



Contents

Volume I **Thesaurus**

Volume II **Index**

Figure 1.1.: The table of contents from *HTE1* (vol.1, p.v).

lexical senses. Section 1.6 discusses two constituents that are found across these main components, namely cross-references and editorial commentaries. Before the conclusion of the chapter, section 1.7 provides an overview of resources that, content-wise, have much in common with historical language thesauri.

1.1. Historical language thesauri

At the time of writing, historical language thesauri are relatively rare.[1] This chapter discusses eight thesauri that cover historical variants of English and Scots, in their various published forms. Only one encompasses the full history of the language which it covers; the others focus on a specific period and, often, a selected domain. A brief overview of the eight thesauri, and the various editions they are available in, is presented below (in order of publication date); Appendix 1.A contains illustrative images of each thesaurus.

[1] Hartmann, *Encyclopedia of Language and Linguistics*, 2nd edn, s.v. 'Thesauruses'; Kay and Alexander, 'Diachronic and Synchronic Thesauruses'.

A Dictionary of Selected Synonyms of the Principal Indo-European Languages (DSSPIEL)
DSSPIEL captures synonyms from a number of Indo-European languages. It includes Greek, Latin, Spanish, Polish, and Russian, amongst others, but also historical forms of languages, such as Old, Middle, and Modern High German (or, as it is known in this thesaurus, New High German). This multi-lingual thesaurus appeared in print in 1949.

The Scots Thesaurus (ScT)
ScT captures the Lowland Scots lexis available throughout history, from its twelfth-century beginnings to the present. It appeared in print in 1990.

A Shakespeare Thesaurus (ShT)
ShT captures the vocabulary employed by the English poet and playwright William Shakespeare (1564-1616) throughout his works. It appeared in print in 1993.

A Thesaurus of Old English (TOE)
TOE captures the lexis of Old English, the variant of English spoken between roughly 500 and 1100 by the Anglo-Saxons. This thesaurus groups synonyms and provides insight into their distribution in the surviving Old English texts. Table 1.1 lists four editions of *TOE* that can be distinguished.

Edition	Type	Description	Published
TOE1	print	the first publication of the thesaurus	1995
TOE2	print	a new impression, containing minor corrections and additions	2000
TOE3	electronic	a website containing further corrections, based on new knowledge that stemmed "largely from completed sections of the Toronto Dictionary of Old English" [2] — *no longer available*	2005
TOE4	electronic	a visual overhaul and relocation of the website of *TOE3* [under rolling revision]	2015

Table 1.1.: Editions of *TOE*

Love, Sex and Marriage (LSM)
LSM captures the English lexis available for love, sex, and marriage throughout the entire history of English. This thesaurus groups synonyms and provides insight into their use in time and place. *LSM* appeared in print in 1999.

Historical Thesaurus of English (HTE)
HTE captures the English lexis that has existed throughout its 1300-year history, from Old English up to Modern English. This thesaurus groups synonyms and provides insight into their use in time and place. Table 1.2 lists three editions of *HTE* that can be distinguished.

Edition	Type	Description	Published
HTE1	print	the first publication of the thesaurus	2009
HTE2	electronic	a digital edition incorporated into *OED Online* [under rolling revision]	2010
HTE3	electronic	a stand-alone website made freely available on the fiftieth anniversary of the founding of the thesaurus project [under rolling revision]	2014

Table 1.2.: Editions of *HTE*

Historical Thesaurus of Scots (HTS)
HTS will capture the vocabulary of Scots throughout its known history. At the time of writing, the thesaurus is still in its pilot phase and focuses on a few selected fields. The pilot version of this thesaurus was published online as a website in 2015.

The Bilingual Thesaurus of Everyday Life in Medieval England (BTH)
BTH contains vocabulary of two languages in use in Medieval England, namely Anglo-French and Middle English, relating to everyday life. This thesaurus groups words and phrases considered synonymous (or possible translations, when dealing with terms from both languages covered) and includes dates of attestation. The thesaurus was published online as a website in 2019.

1.2. Main components

The content of historical language thesauri consists of a number of main components, each with its own constituents, and relations between these components. Werner Hüllen, in his thorough treatment of the history of Roget's *Thesaurus*, distinguishes two main components of thesauri proper: their macrostructure and microstructure.[3] This distinction is common in lexicography and applied more generally also to dictionaries.[4] The term *macrostructure* denotes the arrangement of entries; *microstructure* indicates the structure applied *within* entries. Although this distinction can be applied to historical language thesauri, too, maintaining it is not without its drawbacks.

Firstly, what an entry exactly is can differ from one thesaurus to the next due to differences in presentation. In *LSM*, for instance, an entry consists of a short sense definition (i.e., the name of the category), followed by a list of English words and phrases that, historically, have had that meaning. In *ScT*, an entry is a word with one or more senses. Entries can thus range from multiple sets of synonyms to a single set of these, or even to a single word sense.[5] As a

[3] Hüllen, *A History of Roget's* Thesaurus, pp. 278–84.
[4] Additionally, the entire structure of the publication, along with its introductory apparatus (Foreword, Preface, etc.), is referred to as the megastructure. See Svensén, *A Handbook of Lexicography*, p. 379.
[5] What constitutes an entry and which information is presented as part of an entry during the

consequence, discussions based on macrostructure and microstructure forestall comparisons of thesauri based not on presentation but on what they convey: their informational content. The latter is with which this chapter is concerned.

Secondly, the thesaurus taxonomy can be part of both the macrostructure and microstructure. This is the case for *TOE* and *HTE*. Their entries, as presented in print editions, are groups of synonyms that further expand on, or specialize, the categories found in the overarching macrostructure.[6] Rather than discussing the hierarchy in the macrostructure and microstructure separately, reviewers of the historical language thesauri tend to discuss the topical system in its entirety.[7] Their choice to do so is unsurprising, since which categories of the taxonomy belong to the macrostructure and which to the microstructure can be rather obscure. That is to say, the categories and the items they contain are sometimes presented in a similar manner for both categories belonging to the macrostructure and those considered part of the microstructure. In *TOE2*, for instance, categories found in both structures are presented in bold and are followed by Old English synonyms that express their meaning (see categories "01 Earth, world" and ".As God's creation" in Figure 1.A.7 in Appendix 1.A). The sole distinguishing factor between categories in its macrostructure and microstructure here appears to be that which precedes the category name: a string of numbers, in the case of the macrostructure, or points.[8]

Thirdly, the distinction does not carry over well into electronic environments. Digital editions of historical language thesauri are no longer bound by the limitations that applied to their printed counterparts. Information is dispersed, accessible through hyperlinks, and not presented in a strict sequence and a single visualisation. The electronic editions of *TOE*, *HTE*, and *HTS* focus on another distinction instead: between categories and the words they contain. Their search engines allow users to search amongst the former or the latter. to a singe information proper is separated from its presentation — a design principle that is prevalent in information sciences.[9]

In light of the above, the traditional distinction between macrostructure and microstructure should be abandoned. Instead, thesaurus content demands an analysis through the main logical components identified in digital editions of historical language thesauri. Thus, there are three parts of thesauri: (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 (3) relations of synonymy, indicated through groupings of lexical

editorial process can influence the compilation of the lexicographic work (see, for instance, the practice of employing template entries as discussed in Atkins and Rundell, *The Oxford Guide to Practical Lexicography*, pp. 123-8.

[6] Such fine-grained distinction in meaning between groupings causes the microstructure to be termed distinctive as opposed to cumulative (Kay and Alexander, 'Diachronic and Synchronic Thesauruses', p. 370).

[7] E.g., Coleman, Review of *HTE1*; Diller, Review of *HTE1*; Görlach, Review of *TOE1*; Momma, Review of *TOE2*.

[8] This difference, concerning the identification of categories, is discussed further in section 1.3.3.

[9] Separation of structure and presentation has been advised, for instance, for web pages through the HTML and CSS standards. See W3C, 'HTML & CSS'.

senses.[10] These components, shown in Figure 1.2 (and its legend in Figure 1.3), are discussed separately in the next sections, starting with the topical system.

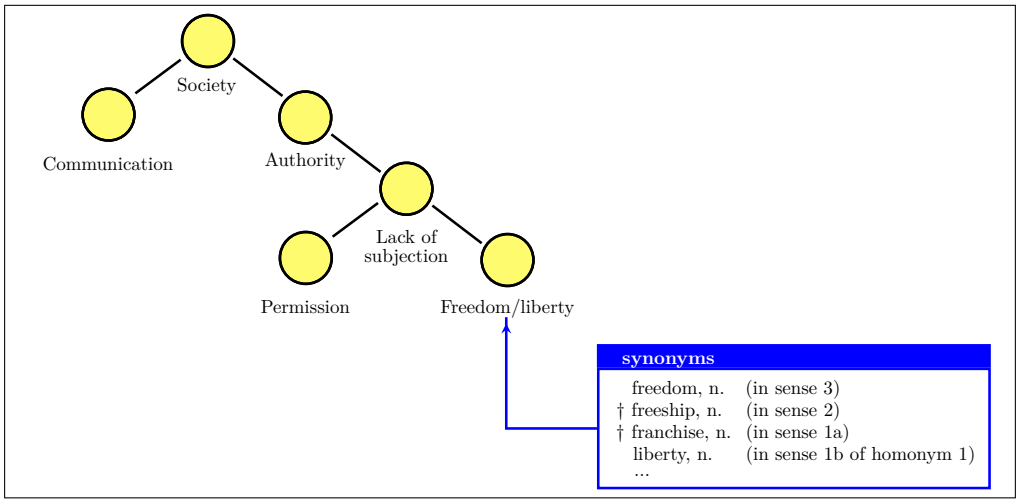


Figure 1.2.: Three logical components of a thesaurus, as found in *HTE2*.

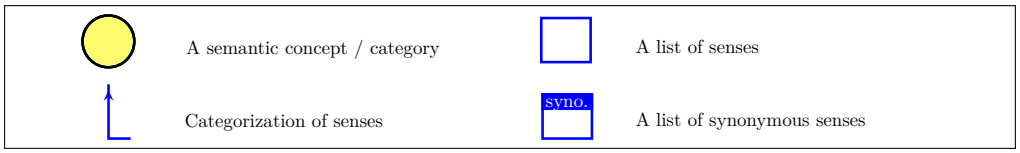


Figure 1.3.: Legend.

1.3. Topical system

Dictionaries commonly employ an alphabetical ordering of their items. Thesauri, in contrast, organize their items according to their meaning through a topical structure. This overarching structure offers users generic meanings as a starting point, which branch out to meanings increasingly specific. Once users locate the meaning that they are interested in, they are presented with the words or phrases that express that meaning. This overarching system thus allows the thesaurus user to move from meaning to lexical item.

Topical systems in thesauri go by different names. Some thesauri, including *LSM*, *DSSPIEL*, and the paper editions of *TOE*, refer to their overarching system as the classification system. Others, including *HTE*, *HTS*, *ScT*, and the electronic edition of *TOE*, position their topical structure as consisting of categories rather than classes, suggesting that their lexis has been categorized

[10] The term *lexical senses* corresponds with the notion of *lexical unit* as defined by Alan Cruse: "the union of a single sense with a lexical form" (*Lexical Semantics*, p. 77). Although relations other than synonymy, such as metonymy and polysemy, may be discerned in thesauri, too, they are, with the exception of *HTE2*, not captured explicitly in the thesauri analysed.

rather than classified. The terms *classification* and *categorization* are often thought to be synonymous and, as a result, "literature on categorization is riddled with passages where the terms [..] are used indiscriminately".[11] In fact, the lack of distinction between *class/classification* and *category/categorization* can be found even in the entry on thesauri in the *Encyclopedia of Language and Linguistics*, which contains a paragraph that speaks of categories in a classification system.[12] Nevertheless, there are differences worth noting.

Although classification and categorization are both means to group items, a classification provides the more rigid structure of the two.[13] The criteria for assigning items to a particular class are predetermined principles that ensure that the classes are mutually-exclusive and non-overlapping. In other words, an item can be assigned to only one specific branch in a classification system. Categories, by contrast, are formed based on context or perceived similarity rather than on predetermined principles. Their boundaries can be "fuzzy": some items may be considered better representative members of a category than others. Which items are assigned to a category and which ones are not therefore remains relatively flexible. Items that are categorized can belong to multiple groups, possibly even located in different branches of a hierarchy (if indeed categories are positioned within a hierarchy). The flexibility that categorization offers over classification is deemed "essential" for historical semantics in particular, according to the editors of *HTE*.[14] As a consequence, the terms *category* and *categorization* appear more suitable for historical language thesauri than *class* and *classification*. The former are therefore used in the remainder of this study.

The logic behind the topical system and its categories tends to be the "central question" for reviewers of thesauri.[15] Manfred Görlach, for instance, has expressed his disappointment in finding in-depth discussion on this matter absent from *TOE1*.[16] Görlach is not alone in criticizing the logic applied by editors (regardless of whether it is made explicit or not) and the resulting topical system.[17] Such criticism did not catch the creators of the discussed thesauri unaware, however.[18] In fact, both creators and reviewers are conscious of the fact that, due to the subjective choices involved in constructing a topical system, it is unlikely that one can be formed that will satisfy every user.[19]

---

[11] Jacob, 'Classification and Categorization', p. 527.

[12] Hartmann, *Encyclopedia of Language and Linguistics*, s.v. 'Thesauruses'.

[13] Jacob, 'Classification and Categorization', pp. 527–31.

[14] *HTE1*, p. xix. The fact that categorized items may differ in how well they reflect the grouping to which they are assigned is in line with what is known as the prototype theory, one of the more recent development in semantics. For an introduction to prototype theory, see Taylor, *Encyclopedia of Language and Linguistics*, s.v. 'Prototype Semantics'.

[15] Diller, Review of *HTE1*, p. 321.

[16] Görlach, Review of *TOE1*, p. 398.

[17] See, for instance, Dance, Review of *TOE1*, p. 312.

[18] In the foreword or afterword of thesauri, editors typically convey their awareness that the fashioned topical system is only a best attempt and may well find disagreement with others (see *DSSPIEL*, p. xiii; *ShT*, p. viii; *TOE2*, pp. xxxv–xxxvi; *HTE1*, p. xix).

[19] For creators, see *DSSPIEL*, p. xiii; *ShT*, pp. viii–ix; *TOE2*, pp. xxxv–xxxvi; *HTE1*, p. xix; *LSM*, p. 22. For reviewers, see Diller, Review of *HTE1*, p. 322; Peters, Review of *LSM*, p. 400.

Andreas Fischer has explained the necessity of subjective choices in fashioning a thesaurus. The first reason, he contends, is that any fashioned topical system will be influenced by the "[c]ultural conceptions and practices" of its creators.[20] The second reason is that "human beings will see most concepts as belonging to several categories" whereas thesauri typically limit the number of categories in order to achieve clarity and brevity.[21] Animals, as Fischer points out, can be allocated in the topical system of a thesaurus under categories of "domestic or wild, or they can be distinguished according to what they eat".[22] Based on these premises, Hans-Jürgen Diller convincingly argues that "[t]here is no one right classification [or categorization]; there are only more and less useful ones".[23] Of course, its usefulness will depend on the context in (or perhaps, rather, the purpose for) which the thesaurus and its topical structure is to be used. The following subsections address three aspects of topical systems: their structure and constituents (i.e., categories), co-ordination of these constituents, and their identification.

1.3.1. Structure of topical systems and their categories

The topical system of a thesaurus, which enables users to move from meaning to lexical items that convey that meaning, consists of categories that are placed in a hierarchical structure. This overarching structure, which is used to organize words and phrases, is not unlike the taxonomies of animals and plants created by the eighteenth-century biologist Carl Linnaeus (1707-1778) and later expanded by the zoologist Georges Cuvier (1769-1832).[24] In these tree-like structures, the most generic or abstract concepts are used as roots, which branch out to concepts increasingly specific in meaning. The concepts lower down in the taxonomy are said to be subordinate to the more abstract ones they branch from.[25] The ones from which they branch, in turn, are their superordinates (see Figure 1.4).

A number of the historical language thesauri distinguish *types* of categories in their topical system.[26] The print editions of *TOE*, for example, distinguish between categories deserving of a numbered identification, and those that do not. To illustrate, the first two categories shown in TOE2, presented in Figure 1.A.7 in Appendix 1.A, are the numbered category "01 Earth, world" and the unnumbered category "As God's creation". Whether a category belongs to the one or the other variant in *TOE* "depends partly on perception of the taxonomy

[20] Fischer, 'The Notional Structure of Thesauruses', p. 55.
[21] Ibid., p. 55; see also Kay, 'When Ignorance is Wisdom', pp. 66–7.
[22] Fischer, 'The Notional Structure of Thesauruses', p. 55.
[23] Diller, Review of *HTE1*, p. 322.
[24] Faria, 'Georges Cuvier et le premier paradigme de la paléontologie'.
[25] Saeed, *Semantics*, p. 68.
[26] Such distinctions are found in historical language thesauri of which their topical system captures fine-grained distinctions as opposed to only coarse ones and are, as a result, labelled distinctive: *LSM*, *TOE*, *HTE*, *HTS*, and *BTH*. All but the latter distinguish types of categories. Chapter 3 describes such distinctions of category types in more detail in section 3.4. Thesauri that do not classify as distinctive are labelled cumulative. See Kay and Alexander, 'Diachronic and Synchronic Thesauruses', p. 370.

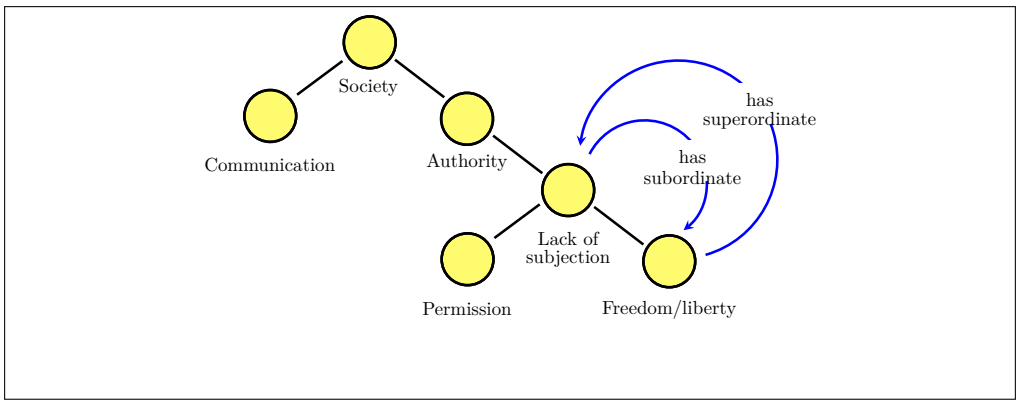


Figure 1.4.: Sample of the topical system of *HTE3*.

and partly on how many words it [the category in question] contains".[27] In other words, there is no true distinction between the different types of categories other than that numbered ones tend to portray larger semantic distances between a superordinate-subordinate pair than unnumbered categories in their vicinity do. Although the current electronic edition of *TOE* supplies these previously unnumbered categories, too, with an identification number (e.g., "01|01 As God's creation"), the two types of categories remain distinguished from each other in presentation and manner of reference (see the separate section of "subcategories" in Figure 1.A.9), in the panel on the right, in Appendix 1.A). Similar distinctions between types of categories are found in *HTE*, *HTS*, and *LSM*. Before detailing the manner in which different types of categories are recognised in historical language thesauri (section 1.3.3), the chapter will discuss the order of co-ordinate categories in the topical system.

1.3.2. Order of co-ordinate categories

A category can have multiple, co-ordinate subordinate categories. In Figure 1.4, for example, the categories "Permission" and "Freedom/liberty" are co-ordinate — both subordinate to the category "Lack of subjection". The order in which such co-ordinate categories are presented in historical language thesauri depends on various factors. To illustrate, *TOE* maintains an order based on meaning, where possible, such as from "Head" to "Tail" for animal parts.[28] If no such order is apparent, the thesaurus resorts to an alphabetical order.[29] *LSM*, similarly, displays an order based on meaning in parts of its topical system. Its Love section "is divided mainly by types of love and degrees of intensity",[30] resulting in categories under "Types of Kissing" to be ordered from "Saluting" and "Pecking" to, at the end of the scale, "French-Kissing". In its section on Marriage, female categories are listed before male ones, as the editor considers that "[i]t would be

[27] *TOE2*, p. xxxiii.
[28] *TOE2*, p. xxxv.
[29] *TOE2*, p. xxxv.
[30] *LSM*, p. 22.

strange, for instance, to list male prostitutes before female, or husbands before wives".[31] Whether this line of thinking is influenced by the cultural conceptions of the past (associated with the lexis in this thesaurus) or those of the editor is left unmentioned. More importantly, in the present context, it is evident that editors may occasionally deem a particular order of co-ordinate categories desirable for users — one that may well be subjective.

Co-ordinate categories in historical language thesauri may not only be ordered based on their meaning but also on grammatical features. *HTE*, *HTS*, and *LSM*, for instance, differentiate between parts of speech in their topical systems. These categories tend to follow a fixed sequence. In *HTE*, the order adhered to is the following: noun, adjective, adverb, verb, phrase, interjection, conjunction, preposition.[32] This order is typically found in *LSM* as well. Here, though, the order may be adjusted in some locations of the topical system. Thus, under "L/01.06 Self-Love", one finds the category "pert. to narcissism/a narcissist" for adjectives preceding "falling in love with sme/sth of one's own creation" for nouns. The editor of *LSM* states that any deviations found in the thesaurus are "based on the earliest citation date, the relative number of terms for each part of speech, and the ease of defining words in terms of one other".[33] For categories, too, then, editors may on occasion desire to impose an ordering other than the default.

1.3.3. Identification of categories

Categories in historical language thesauri are typically identified through a name preceded by a number or string that codifies that category's location in the thesaurus. The exact identification format employed for a category depends largely on the *type* of category and its position in the topical system, as will become evident in this section. Table 1.3 illustrates the identification system employed in many of the historical language thesauri with an example taken from *LSM*.

L Love L.03 Friend L.03.02 Companion	*Category type I*
L.03.02/01 Types of Companion L.03.02/01.01 Travelling Companion	*Category type II*
a travelling companion .pert. to a travelling companion	*Category type III*
adj adv	*Category type IV / PoS*

Table 1.3.: Hierarchy and category types of adjectives in *LSM* "pert[aining] to a travelling companion".

[31] *LSM*, p. 23.
[32] *HTE1*, p. xxii.
[33] *LSM*, p. 26.

As Table 1.3 shows, *LSM* distinguishes four types of categories. Categories of type I are identified by a name and a string that codifies the location in the hierarchy. For example, the string "L.03.02" used to identify the "Companion" category indicates that this category is the second one subordinate to the third ("Friend") of the top category L ("Love").[34]

Categories of type II in *LSM* are indicated by a similar system, but a forward slash in the string (separating categories of type I and type II) indicates that these subordinate categories are semantically closer to their superordinate category than subordinate categories of type I. For this reason of semantic proximity, the printed thesaurus presents categories of type II (e.g., "L.03.02/01 Types of Companion") directly after their superordinate category (e.g., "L.03.02 Companion"). Subordinate categories of type I (e.g., "L.03.02.01 Acquaintance") follow after those of type II.

Besides categories identified in *LSM* via a string codifying their exact location in the hierarchy, the thesaurus also contains categories that are not identified in such a manner. *LSM* categories of type III (in Table 1.3 by the names of "a travelling companion" and "pert. to a travelling companion") are identified only by a name or sense definition. To indicate hierarchical relations between categories of this type, *LSM* employs a "system of indented points".[35] A category name visually preceded by a point is subordinate to the first category listed above it that is preceded by one fewer. In Table 1.3, the category "pert. to a travelling companion", which is preceded by a single point, is subordinate to the category "a travelling companion", which is preceded by no point at all.[36] Lastly, *LSM* type IV categories state the part of speech.

The number of category types, and the identification system in place, may differ from thesaurus to thesaurus. *LSM* is certainly not the only historical language thesaurus that codifies the location of a category in a string. In fact, all historical language thesauri covered in the chapter employ such a system. This practice suggests that users are thought to benefit from references to categories that, to some degree, include information on the context of the topical system. Additionally, all of the historical language thesauri that distinguish types of categories convey these distinctions in the codified string through special delimiters (such as a forward slash). Although these identification systems emphasize the hierarchical structure, they are not without drawbacks. The resulting identification strings can be rather lengthy — especially in large, distinctive thesauri, such as *HTE*. As a result, reviewers have dubbed these strings "off-putting", "unwieldy", and "impossible to hold in one's head for more than a few seconds without re-checking in the text".[37] Of course, this last remark may indicate that the criticism holds more so for print editions of thesauri rather

[34] *LSM*, p. 21.

[35] *LSM*, p. 26.

[36] The editor of *LSM* considers categories of type III not to be categories in the topical system, but instead sets of synonyms identified by a definition positioned within such categories (*LSM*, pp. 25–6). The hierarchy between these groups, however, form a taxonomy based on meaning, characteristic of categories within a topical system.

[37] Coleman, Review of *HTE1*, p. 206; Görlach, Review of *HTE1*, p. 195; Brewer, Review of *HTE1*, p. 804.

than digital ones, in which it is not necessary to remember the entire string, as it can be copied, bookmarked, or provided as hyperlink to the corresponding section in the thesaurus.

Use of indented points in the identification system, such as with category type III in *LSM*, is also found within the print editions of *TOE*. There, indented points are used with the second of its two category types. This system has been replaced in its electronic editions, which favour an identification system that codifies the precise location in a string for both its category types (see Table 1.4). Although these newer *TOE* editions do not clarify the reasoning behind the change, one may assume its intention has been to unify the referencing system for all categories and that, in the digital editions, there is more visual space available than in the paper editions to present numbers for all categories.

12 (n.) Power, might 12.01 (n.) Power, control, sway 12.01.01 (n.) Authority 12.01.01.10 (n.) Freedom, being free	*Category type I*
12.01.01.10|07 (adj.) Free, not in bondage 12.01.01.10|07.01 (adj.) Free to go	*Category type II*

Table 1.4.: Hierarchy and category types of *TOE4* "Free to go".

In addition to differences in their numbered identification, thesauri can also differentiate in how they incorporate part of speech. Whereas in *LSM*, type IV categories state the part of speech only, other historical language thesauri may integrate this information in the name of the category, too. This practice is found in both *HTE* and *HTS*. Thus, we find in *HTE1* the categories "01 (n) The world" and "01 (adv) Earthly", which are the nominal and adverbial part of speech categories subordinate to the first top category of the thesaurus. Moreover, these thesauri show that part of speech categories do not necessarily need to be positioned as the lowest level of categories for which no further subordination is possible: Both *HTE* and *HTS* employ part of speech categories as their second type of category and allow subordination by a third type of category (see Table 1.5).[38]

To summarize, historical language thesauri identify categories through a system that incorporates their position in the topical system. The exact identification mechanism can differ between thesauri. Those thesauri that distinguish multiple types of categories foreground these differences in the identification system (e.g., through a forward slash or pipe separating two types of categories). For thesauri in which all lexical items positioned at a category share their part of speech, this property may be presented at the category instead

[38] It may be worth noting that *HTE3* appears to position its part of speech identification in what could be considered an odd position. The category on nylon, for instance, is said to require citing through the reference "01.08.01.09|01.03 (n) Man-made textiles :: synthetic :: nylon". Considering that the part of speech category is located earlier in the classification system though, something alike the following would be more accurate (albeit by some perhaps considered visually less attractive) for a codified location string: "01.08.01.09 (n)|01.03".

03 Society 03.04 Authority 03.04.01 Power	*Category type I*
03.04.01 (adj.) Powerful/mighty	*Category type II*
03.04.01|03 (adj.) All-powerful 03.04.01|03.01 (adj.) Jointly	*Category type III*

Table 1.5.: Hierarchy and category types of *HTE3* "Jointly (all-powerful)".

of at the lexical sense, reflected by the category name, and incorporated in the identification string.

1.4. Lexical senses

The topical system of a thesaurus categorizes elements referred to as lexical senses. These elements are more specific in nature than words or phrases, known as lexemes. By drawing on an example from *ShT*, this section aims to clarify what lexical senses are and that it is indeed these elements, and not lexemes, that are categorized. Consider the *ShT* category "01.02 sky", under which one finds the following item printed:[39]

> heaven, n.

The head-form "heaven" in this example is similar in appearance to a headword, or lemma, found in typical dictionaries, which suggests that lexemes are categorized by thesauri. The fictitious dictionary entry below, however, demonstrates otherwise.

> **heaven**, n. 1) abode of one or more gods 2) the sky

It is evident that the "heaven, n." entry in *ShT*, found in the category "01.02 sky", represents the lexeme *heaven* in not all of its senses listed above but in only the second sense. Werner Hüllen, who has thoroughly researched the topical tradition of thesauri, acknowledges that items in thesauri indeed represent senses of lexemes rather than lexemes in their entirety:

> Strictly speaking, topical dictionaries [i.e., thesauri] have no headwords but head-forms as linguistic dummies for their meanings [i.e., senses]. Admittedly, a highly developed linguistic awareness is needed to keep this difference in mind when using a topical dictionary. Hence the humorous criticism that in order to work with such dictionaries you must be so highly educated that you do not need to consult a dictionary at all.[40]

[39] Information on lexical senses or their lexemes, such as the part of speech, may be presented at the category instead (or posited as a category of its own) if the information applies to all items in that category. This condition is not met in *ShT* for the part of speech of lexical senses positioned at a given category.

[40] Hüllen, *English Dictionaries, 800-1700*, p. 13.

This claim is strengthened by the fact that the type of synonymy indicated by thesauri is a relation between senses and not between lexemes or word-forms, as will become clear in section 1.5.

Confirmation that thesauri categorize lexical senses can be found in the electronic edition of *HTE* that has been incorporated into *OED Online* (i.e., *HTE2*). This particular edition of *HTE* takes advantage of both the topical structure of the thesaurus and the full dictionary entries of the *OED*. This set-up – unique amongst the historical language thesauri analysed – allows for a closer investigation of the relation between a thesaurus and entries in a dictionary. Dictionary entries in *OED Online* refer to thesaurus categories at each lexical sense separately (see Figure 1.5). Conversely, the thesaurus categories in this edition list the senses they contain and provide hyperlinks not simply to dictionary entries but to specific senses *within* these entries (see Figure 1.6). As such, it is evident that *HTE* indeed categorizes senses of lexemes, and not lexemes as a whole.[41] The following subsections discuss the structure, co-ordination, and identification of these lexical senses in historical language thesauri.

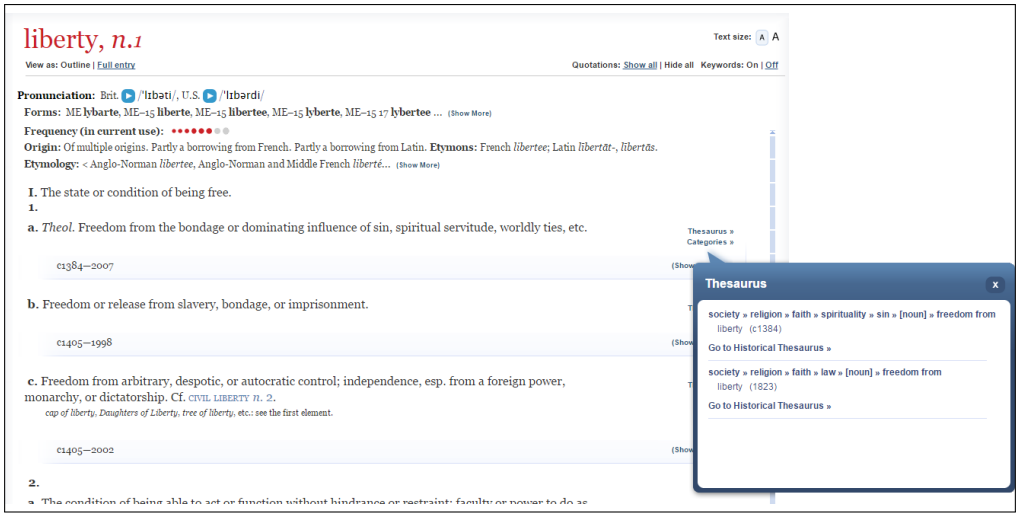


Figure 1.5.: Entry 'liberty, n.1' in *OED Online* (3rd edn). A pop-up window, on the right, indicates that sense 1a is assigned to two categories of the historical language thesaurus *HTE2*.

1.4.1. Structure of lexical senses

For any lexicographical work, including thesauri, the question of which information is provided per lexical sense depends on the audience and purpose of the work as its editors perceived it.[42] The structure of senses in ordinary dictionaries, for instance, typically includes the part of speech, definition, and

[41] This conclusion is supported by the fact that the *OED* labels usage features per sense, and that *HTE* has picked up these labels for its categorized lexis, presenting these at each head-form in both the paper and stand-alone electronic edition. The head-forms in this thesaurus, therefore, must indeed represent senses.

[42] Atikins and Rundell, *The Oxford Guide to Practical Lexicography*, p. 200.

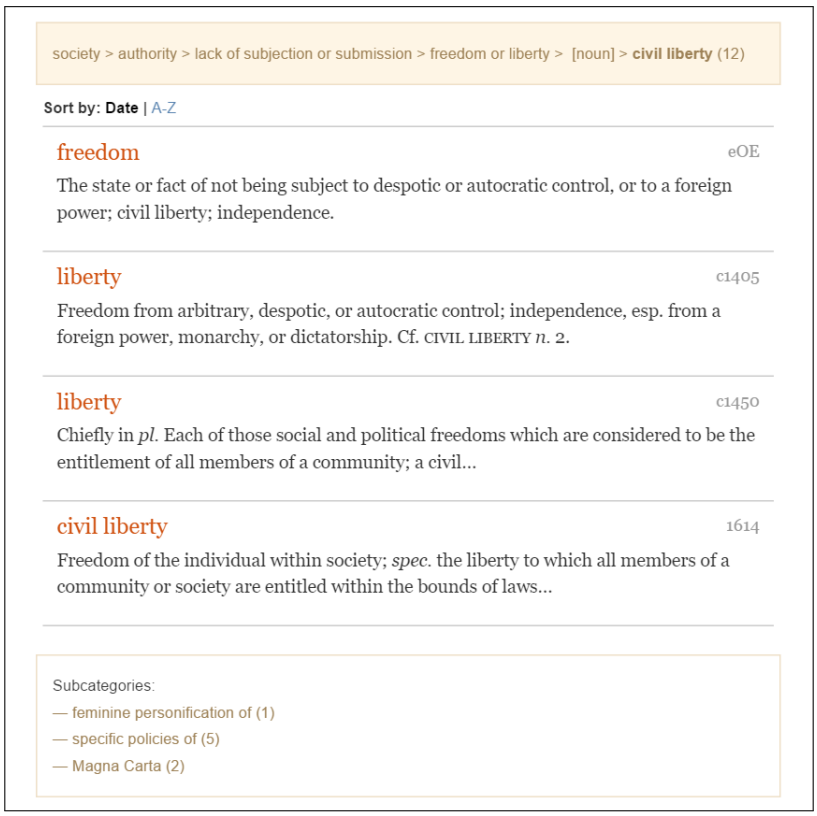


Figure 1.6.: The category "civil liberty" in *HTE2*. Two senses of *liberty* are included in this category. The first of these links to sense 1c of *OED Online* entry 'liberty, n.1'; the second to sense 2c.

possibly a number of other components, such as quotations and meta language indicating register or regional use.[43] These components aid in clarifying the unknown: the meaning and usage of a certain word or phrase. For thesauri, the unknown is not the meaning, which is captured in the topical system, but the words or phrases that express that meaning. Hence thesauri often omit definitions for their senses; information on their meaning is already indicated (to a certain degree) through their placement in the overarching hierarchy.[44] Moreover, any additional information per lexical sense, such as etymology or register, may already be captured in other reference works and is sometimes considered superfluous in thesauri.[45] Nevertheless, the historical language thesauri analysed in the chapter all include such additional information (such as part of speech in *ShT* and use restricted to poetic texts in *TOE*), which may be attributed to their purpose of acquainting users with words and phrases from a historical context rather than a contemporary one with which they may be more familiar. What additional information is present, and in which of these thesauri, will be discussed shortly.

[43] Ibid., pp. 200-57.
[44] *TOE2*, p.xxxiii; *HTE1*, p. xxii.
[45] *ShT*, p. x; *ScT*, p. xv.

The order of components presented for a lexical sense typically adheres to a strict sequence in lexicographic works.[46] The order decided upon by editors, which may vary between such works, is mainly intended to present the user of the resource with information in a structured and consistent manner. Through such a sequence, and through visual clues to distinguish the components, users are aided in recognising the information presented and in locating the elements in which they are interested. Although each historical language thesaurus edition may adhere to its own sequence of components, the works analysed here tend to start with a form of the lexical item (its head-form), followed by the part of speech and, if such components are included, the definition, usage features, and lastly external references. These components are discussed, separately, in the following subsections.

Part of speech

The majority of the historical language thesauri analysed explicitly state the part of speech for their lexical senses. This information is presented either per sense separately (*ShT*, *ScT*, *BTH*) or, more often, per category or set that groups lexical senses (*LSM*, *HTS*, all editions of *HTE*, and the electronic editions of *TOE*). The remaining thesauri leave it to the user to infer the part of speech of lexical senses from the meaning indicated by the names of the categories (*DSSPIEL*, print editions of *TOE*). In these last-mentioned thesauri, the part of speech is explicitly indicated only under special circumstances, such as when the editor considers the part of speech to be unclear from the context. To illustrate, words located under *DSSPIEL* category "1.85 *Burn*" are explicitly labelled as verbs in order to ensure they are not interpreted as nouns. It should be noted that the indication of a part of speech for a lexical sense also holds, of course, for the lexeme to which that sense is attributed.

Definition

A definition for lexical senses is another component found amongst the historical language thesauri of Scots and English. As thesauri already indicate the meaning of lexical senses through their placement in the topical system, definitions are often omitted in the structure of the lexical sense. In fact, the only thesauri analysed that contain such definitions for each of their lexical senses (again, beyond the definitions that can be construed using the topical system) are those covering the Scottish lexis: *ScT* and *HTS*. In her review of *ScT*, Betty Kirkpatrick explains the need for the definitions for these particular thesauri:

> There are innovations in the Scots Thesaurus. Unlike other thesauruses it has definitions. This is not only essential for people from outside Scotland but for many Scots living in Scotland who need to be acquainted or re-acquainted with their native tongue.[47]

[46] Atkins and Rundell, *The Oxford Guide to Practical Lexicography*, pp. 200–57.
[47] Kirkpatrick, Review of *ScT*, p. 306.

Here, therefore, the inclusion of definitions is warranted by the intended audience for these thesauri. The need to know what a word exactly means is deemed important enough for its inclusion as opposed to assuming that users are already aware of the distinctions between related or synonymous words, are not interested in such information, or are willing to consult other reference works. For the majority of the historical language thesauri analysed in this chapter, users are indeed encouraged to consult other sources for exact definitions (as well as for other components) when the need for such information arises.[48]

Language

The language of a lexical sense is indicated explicitly in only two of the historical language thesauri analysed. The component is found in *DSSPIEL* and *BTH*, multilingual thesauri fashioned specifically for the purpose of contrasting words from multiple languages that denote the same concept (Indo-European languages, in the case of *DSSPIEL*; Anglo-Norman and Middle English in *BTH*). *ShT*, too, may be considered a multilingual thesaurus, as its editor has opted not to confine the thesaurus to Shakespeare's attested English lexis only. The "foreign words" found in Shakespeare's works, far fewer in number than the English ones, "are normally those in foreign-language contexts only".[49] That is to say, characters in a play may be French or Italian and converse in their native tongues. The play *Henry V*, for instance, contains a scene in which the French princess Catherine tries to learn a number of English words from her maid.[50] The conversation is performed entirely in French, apart from those words that are being taught, of course. To indicate this distinction in language in Shakespearian vocabulary, *ShT* marks foreign words with a label ('lat.' for Latin, 'fr.' for French, and 'it.' for Italian). The other thesauri analysed deal with the lexis from what is considered a single language, albeit with varying dialects or variants. An explicit reminder of the language at every lexical sense contained within is, understandably, deemed unwarranted by their editors.

Usage features

Information on usage and distribution is, according to the editor of *LSM*, "not normally included" in the structure of lexical senses in thesauri.[51] In contrast, historical language thesauri appear to be an exception to this rule: five out of the eight analysed include information on usage features. One must conclude, therefore, that users of historical language thesauri are thought to require, or rather desire, more information on the included lexical senses than users of most contemporary thesauri do. Of course, historical lexis may take on forms that are

[48] *ShT* dubs existing dictionaries "indispensible companions to the present work" (p. x). *HTE* encourages the user to "return to the *OED* and gather fresh information" on lexical senses (*HTE1*, p. xiv). *TOE* informs its users of the general need for dictionaries next to thesauri: "Compared with a dictionary, any thesaurus is somewhat of a blunt instrument, sacrificing semantic or grammatical specificity to breadth of conceptual coverage." (*TOE2*, p. xxxv).

[49] *ShT*, p. xiii.

[50] Shakespeare, *Henry V*, pp. 140–2.

[51] *LSM*, p. 5.

significantly different from contemporary ones or may have undergone significant changes in their meaning or use — changes that are worth pointing out to users of historical language thesauri through, amongst others, indications of usage features.

Appendix 1.B provides an overview of which kinds of usage information are conveyed systematically in the historical language thesauri analysed. These include indications of restricted use of a given word or phrase to a specific genre, region, dialect, or medium. Most of these features are indicated through labels (e.g., 'poetic' for words specific to poetic diction). In contrast, diachronic usage features are captured in *LSM* and *HTE* through dates of currency instead. Here, a label is used solely for the historical period of Old English (abbreviated to 'OE'), since exact dating of texts and language use for this early medieval period is not as straightforward as for later periods of English.

External reference

The final component found in historical language thesauri of Scots and English is an external reference. In fact, the majority of the historical language thesauri advise the user to pick up additional reference works for further or more detailed information on their included lexis. Many of these thesauri have taken dictionaries as the source of their information, which entails that references to the source material can be maintained during the editorial process. In cases where not all information from a source dictionary is transferred to the resulting thesaurus, references to the source material can offer users valuable insights beyond information the thesaurus editors deem to warrant including directly. To illustrate, *HTE* abandons indications of etymology found in the *OED*.

The form in which external references are made differs between print and electronic editions of the thesauri analysed. Print editions typically refer the user to a source dictionary in the introduction. Providing an external reference per lexical sense instead would offer little to no additional benefit, as it requires the reader to manually access the contents of the external reference work and locate the sense in question. Digital editions of thesauri, in contrast, are capable of facilitating the user in this process. In these editions, references can be offered by hyperlinks that lead the user to specific locations in other digital bodies. Such efficient links are available in the electronic editions of *TOE*, *HTE*, and *HTS*. Hyperlinks per lexical sense refer the user in these cases to either the exact corresponding item in the source dictionary (with *HTS* and *HTE2*) or to a search for the item using the search engine of an external digital body (with *TOE4* and *HTE3*). Of course, external references need not only lead to dictionaries.

TOE4 offers the user external references to University of Toronto's *Dictionary of Old English Web Corpus*. These references allow the user to automatically search for attestations of a word or phrase from *TOE* (with that particular spelling) amongst Old English texts that have come down to us, effectively offering insights into the contexts in which the item is known to have been used. Having discussed this final component and thereby the structure of lexical senses

in their entirety, I turn to other aspects of these items found in historical language thesauri: their order when co-ordinated and, afterwards, their identification.

1.4.2. Order of co-ordinate lexical senses

Co-ordinate lexical senses in historical language thesauri, i.e., senses located at the same thesaurus category, are displayed in a systematic order. This order, as Kay and Alexander observe, appears to depend on the availability of diachronic usage information:

> Within the macrostructure, historical language thesauruses which regard their data as belonging to a single period will usually display synonyms in alphabetically organized lists. Those with a diachronic spread will order lists chronologically, or will compromise by including some information about dates of use within an alphabetical list, as *The Scots Thesaurus* (1990) does.[52]

Indeed, the historical language thesauri that include diachronic usage information in a detailed manner (i.e., *LSM* and *HTE*) order their senses chronologically.[53] *ScT*, as noted, employs an alphabetical ordering instead. Diachronic usage information in *ScT* is rather limited, effectively dividing senses into those that are considered obsolete and those that are not. An ordering based on this rather coarse distinction would still require co-ordinate lexical senses to be organized within these two groups. In lieu of a better alternative, the otherwise meaningless alphabetic ordering can always be employed to order senses. In fact, it should be noted that both *LSM* and *HTE* indeed fall back on alphabetical ordering for co-ordinate senses that have identical diachronic usage information.[54]

Although most historical language thesauri without diachronic usage information employ an alphabetical ordering of co-ordinate lexical senses, not all of them can be said to do so. An alphabetical ordering is applied in *ScT*, *ShT*, *TOE*, and *HTS*.[55] These are thesauri that capture the lexis of a single language or, in the case of *ShT*, at least for the majority of its items. In contrast, the multilingual thesaurus *DSSPIEL* presents the items in three columns based on their language.[56] The first column displays the synonyms for a particular concept taken from Hellenic languages, then those from Italic languages, followed by Celtic ones. The second column displays the lexical senses taken from Germanic

[52] Kay and Alexander, 'Diachronic and Synchronic Thesauruses', p. 372.

[53] *LSM*, p. 15; *HTE1*, p. ix.

[54] *LSM*, p. 28; *HTE1*, p. xxiii.

[55] It should be noted that the alphabetical ordering might be on a particular form of the categorized items. In *TOE3*, for instance, the category "01|01 (n) Earth, world :: As God's creation" contains *gesceaft*, which is found after *sǽ* and *eorþe*. The reason for this ordering is that some Old English words are found in texts both with and without the *ge-* prefix, and that therefore some lexicographical bodies ignore this prefix in their alphabetical ordering (see, for instance, *DOE*).

[56] *DSSPIEL* does not make this order or the reason behind it explicit, suggesting the order is considered logical enough to the readers interested in the comparative linguistics with which the book is concerned.

languages. The third and last column those of Balto-Slavic and afterwards Indo-Iranian languages. Within these groups, too, the order of the languages is set. For the Germanic ones, for instance, Gothic is followed by Old Norse, Danish, Swedish, Old English, Middle English, New English, Dutch, Old High German, Middle High German, and lastly, New German.[57] Here it is possible to see that the order is again based on more than simply chance or the alphabet, as younger variants of a language are preceded by older ones (e.g., New English is preceded by Old and Middle English). In short, the order of lexical senses in a historical language thesaurus can depend on any piece of information associated with them — not limited to diachronic information. This finding suggests that other information, too, such as diatopic usage information, could be used to order lexical senses in historical language thesauri yet to be developed.

As the above discussion has shown, the order of co-ordinate lexical senses is based on explicit information captured for each sense. Any ordering found in the historical language thesauri analysed does not present new information on the lexical senses or provides a better understanding of how these co-ordinate items relate to one another. Instead, the applied order ensures that users can expect to find an item in a position governed by a system that is – or rather, should be – easily grasped.

1.4.3. Identification of lexical senses

Dictionaries and thesauri, both lexicographic works, tend to employ a different identification mechanism for their lexical senses. The dictionary *OED Online*, for instance, contains the following senses for the adjective *politely*.

> **politely,** ***adv.*** **†1.** Smoothly. Obs. 1598–1730 **2.** In a polished or refined manner; elegantly. Now rare. 1624–1868 **3**. Courteously; with good manners; with consideration for the feelings of others. 1748–1993 [58]

A reference to one of these three senses typically includes both the headword and the code that identifies the sense within this particular grouping, for example: *OED Online*, 3rd edn, s.v. 'politely, adv.', sense 3. In the case of thesauri, this practice can only be applied effectively if they maintain a relation between lexical senses and their lexemes. Only *HTE2*, out of the eight historical language thesauri, does so. Consequentially, lexical senses in thesauri often have to be identified in a manner different from that found in typical dictionaries and utilize the topical system in which senses are positioned.

Lexical senses in a thesaurus are spread across its topical system. Senses that in a dictionary would have been grouped under the same headword may well be found at different locations in a thesaurus. The head-form that identifies such a sense, acting for it as a "linguistic dummy", is therefore useful for clear identification only in the context of the category in which that sense is found. References that would include only the head-form of a lexical sense tend to

[57] These are the names for the languages as adopted by *DSSPIEL* (see pp. 1–2).
[58] *OED Online*, 3rd edn, s.v. 'politely, adv.'.

be ambiguous as a result. Referring to "light", for instance, is not sufficient to determine whether its sense categorized as "lamp" is meant or that as "relatively small in weight". The position of a lexical sense in the topical system is therefore, in the majority of historical language thesauri, highly relevant for proper identification. To illustrate, the sense of *franchise* as shown in Figure 1.2 could be identified by referring to "*franchise* as found in 03.04.10.03 n. Freedom/liberty". (The identification and referencing to categories in the topical system has been treated earlier, in section 1.3.3).

In some of the historical language thesauri, the head-form along with the location in the topical system is insufficient to refer to a single sense unambiguously. The *ScT* category "15.6.13 Anger", for instance, contains the following entry:

> **scunner 1** a nuisance. **2** a troublesome person.

Here, the grouping of senses is similar to that found in typical dictionaries. Hence there is the need for including the number that identifies a specific sense within the grouping, too, in references to a lexical sense. The location in the topical system remains relevant even in cases like *ScT*, which appear at first glance to resemble a dictionary in terms of its identification and grouping system. The groupings of senses based on head-form and number are guaranteed to be unique only per category. Thus, one finds the head-form "able" in *ShT* without any further sense division in both category "9.1.1 General", in the sense of being physically fit, and in "10.4 Eating and drinking", in the sense of having an appetite. Only their definition and locations in the topical system can be used to distinguish the two senses. In short, the identification of lexical senses in a thesaurus typically consists of a head-form and perhaps further identification within a group. However, these identifications are not necessarily unique throughout the thesaurus and references should therefore include the position of such senses in the overarching topical system.

1.5. Synonymy

The overarching topical system of a thesaurus provides a structure for words and phrases in their various senses. Senses grouped by thesaurus categories are, per definition, sets with a similar or related meaning. With most historical language thesauri this chapter covers – but not all – such sets indicate an even stronger semantic tie: one of synonymy. The words *start*, *begin*, and *commence*, for instance, can be considered synonymous, since they denote the same concept. In which cases words or phrases are deemed synonymous depends on the definition of the relation to which is adhered.

The semantic relation of synonymy has four prominent definitions in linguistics. Arranged from most to least restrictive, these various types of synonymy are called absolute, perfect, cognitive, and near-synonymy (see Table 1.6). The most restrictive of these, absolute synonymy, requires lexemes to have the same meaning and use for all their senses. Whether this strictest variant of synonymy truly exists is debatable. Words or phrases may not only differ in the

meanings they carry but also in some other aspects, such as their connotation, register, or frequency in use.[59] The other three definitions consider synonymy to be a relation between lexical senses instead of lexemes in all their senses. Perfect synonymy of two lexical senses requires them to be the same in meaning and use regardless of the context in which they are used. Cognitive synonymy stipulates that senses need to be the same in meaning apart from any possible variation in usage between them. The least restrictive definition, that of near-synonymy, calls for a similarity of meaning and requires two senses to be substitutable only in specific contexts.

Relation	**Between**	**Example**
Absolute synonymy	lexemes	*anyone=anybody*
Perfect synonymy	lexical senses	*crucifer=brassica* ('cabbage')
Cognitive synonymy	lexical senses	*peritonsillar abscess=quinsy* ('illness')
Near-synonymy	lexical senses	*to pacify=to placate* ('to calm')

Table 1.6.: Examples of different definitions of synonymy.[60]

Not every thesaurus indicates relations of synonymy according to even the most forgiving of definitions. *ShT* and *ScT*, for instance, omit such indications. To illustrate, *ScT* contains the category "6.2.3 Shipping, navigation" in which the noun *spulyie* and the verb *stell* are adjacent entries. The sense of the former is indicated as "jetsam, anything cast ashore"; the latter as "load (a ship) evenly, trim the cargo in (a ship)". These two senses can hardly be considered interchangeable; even the parts of speech of the items differ.[61] Adjacency here, therefore, does not indicate synonymy. Similarly, *ShT* positions the verb *overblow* and the noun *sea-storm* alongside each other in category "01.06 Wind, storm". These, too, can not be considered synonymous according to any of the four established definitions. In short, grouped items presented at the same location in the topical system of these two thesauri do not indicate anything beyond belonging to the same semantic field.

Although not found in *ShT* and *ScT*, sets of synonyms are present in all the other historical language thesauri analysed: *DSSPIEL*, *TOE*, *LSM*, *HTE*, *HTS*, and *BTH*. The degree to which members of such a set are considered synonyms may vary across these thesauri. *TOE*, for instance, asserts that its grouped items, although they share "at least one component of meaning", are to be seen as loosely synonymous only.[62]

> No claim is made that the words assigned to each group are synonyms in any strict sense of the term, i.e. that they are

[59] Murphy, *Encyclopedia of Language and Linguistics*, s.v. 'Synonymy'.

[60] Based on Murphy, 'Meaning Relations in Dictionaries', pp. 447–8.

[61] Synonymous lexical items are typically thought, in order to be interchangeable, to require a shared part of speech (Murphy, 'Meaning Relations in Dictionaries', p. 439). Be that as it may, some scholars "claim that synonymy is possible between words belonging to different parts of speech (as between the verb *sleeping* and adjective *asleep*)" (Stanojević, 'Cognitive Synonymy: A General Overview', p. 194.

[62] *TOE2*, p. xxxiii.

> mutually substitutable in all or most contexts. Rather, they are loosely synonymous terms which express the concept defined by the headings, which will itself often be a descriptive phrase rather than a single word.[63]

The introduction in *LSM* contains a similar statement:

> Grouping terms together in a thesaurus, even in a thesaurus as detailed as this, does not imply absolute synonymy. Many scholars doubt whether absolute interchangeability is actually possible [...].[64]

In light of these statements, it must come as no surprise that the least restrictive form of synonymy, near-synonymy, has been called "the staple of thesauruses".[65] Lexical senses grouped as synonyms – or rather, as near-synonyms – therefore possess a certain degree of interchangeability.

The presence of synonymy is, in editions of the historical language thesauri analysed, not presented differently from the absence of synonymy (or, to be more precise, from omissions of their presence). To illustrate, all senses in *HTE1* positioned in the same category are considered synonyms; those in *ShT* are not claimed to be (and often are not).[66] In thesauri that capture synonymy, therefore, categories can act as sets of synonyms as well as semantic fields. Even so, there are notable differences between the two of which users should be aware.

The first and perhaps most obvious difference between synonym sets and categories representing a semantic field is that the former imply that the relation between its grouped members is one of interchangeability in certain contexts, whereas this is not necessarily the case for members belonging to a category within any given thesaurus. (As mentioned, this lack of interchangeability for grouped items is found in *ShT* and *ScT*.)

The second difference between synonym sets and categories concerns the membership of lexical senses. Lexical senses located at a particular category are members of that category and any superordinate categories, which can be viewed as generalizations of the semantic field.[67] If these same senses are known to be synonymous, that is, part of a synonym set, they are *not* also synonymous with the senses found in those superordinate categories. In Figure 1.7, for example, the lexical sense of *freedom* at category "Freedom/liberty" is not synonymous with *command* and the other senses listed at "Authority". This sense of *freedom* is, however, categorized as belonging to categories "Society", "Authority", "Lack of subjection", and "Freedom/liberty". Simply put, the placement of a lexical sense in a thesaurus may posit it as part of a specific synonym set and typically signifies the sense belongs to multiple categories — one that is the most specific and others that are superordinate. These differences, as shown for *HTE* in the previous paragraph, tend to be left implicit in the presentation of the thesauri,

[63] *TOE2*, p. xxxiii.
[64] *LSM*, p.15. Coleman refers to publications by Ullman, Firth, Harris, and Baldinger in which doubts about absolute synonymy are raised.
[65] Murphy, 'Meaning Relations in Dictionaries', p. 448.
[66] *HTE1*, p. xxv.
[67] *HTE1*, p. xviii.

and may be left implicit in the way the lexemes and their senses are stored in the datasets behind electronic editions, too.[68]

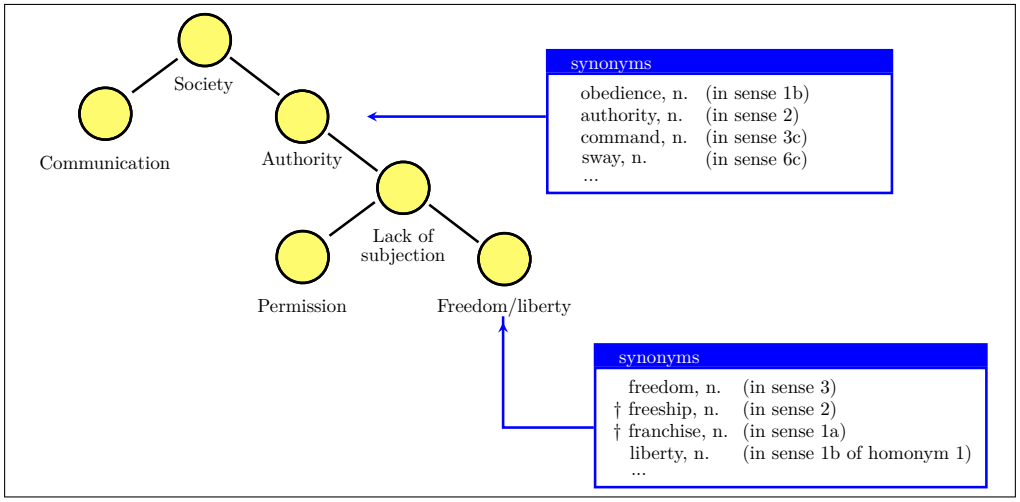


Figure 1.7.: Membership of lexical senses.

1.6. Constituents found across multiple components

Two more constituents, in addition to those described in the previous sections, can be found in the thesauri analysed: cross-references and editorial commentaries. Since these constituents are not confined to a single main component of the thesauri analysed, but occur with both the topical system of historical language thesauri and their lexical senses, they are discussed separately in this section, starting with cross-references.

1.6.1. Cross-references

A number of historical language thesauri contain cross-references in their topical systems, acting as guides to other locations of the taxonomy that are related or deemed relevant. In *TOE4*, for instance, the category "05.11.03 n Period of time, era, epoch" contains two cross-references: one to category "02.01.02.02.02 n Lifetime" and another to "02.01.04 n Age". *LSM* and *ScT*, too, contain cross-references leading the user from one category to another.[69] The need for such cross-references in thesauri appears to arise from subjectivity and ambiguity that are inherent in the act of classification and categorization. Julie Coleman touches upon exactly this matter in her review of *HTE*.

> Boundaries between meanings are not impermeable or permanent. Although the compilers of a historical dictionary will tend to discard

[68] A case in point is *TOE4*, of which the database leaves relations of synonymy implicit (see Chapter 6).

[69] Cross-references are, for instance, found in *ScT* at category "5.2.3 Rain, mist, snow, frost" and in *LSM* at "L/01.05 Family love".

> ambiguous citations if clearer examples are available, it is undeniable that grey areas exist where it is unclear whether sense 1c or 1d is intended. [...] Roget's son went some way towards dealing with this problem [i.e., of needing to place an item in multiple locations in the thesaurus] by expanding the use of cross-references between sections.[70]

The importance of cross-references has also been pointed out in reviews of historical language thesauri that lack them, which mention that their addition would be "very useful" to users.[71]

Cross-references may exist between categories, but they may also be found – and deemed useful – between lexical senses. One such an example could be between polysemous senses of the same lexeme. As Haruko Momma points out, these cross-references are facilitated in print editions of historical language thesauri through their index. Their mention of headwords, along with the thesaurus categories that they are located in, allows one to look up metaphorical meanings.

> For example, the first volume [of *TOE* in print, containing the thesaurus proper] shows that Old English has three synonyms for 'hand', and the second volume [containing the index] shows that *folm* has the literal meaning alone, whereas *hand* and *mund* have metaphorical meanings each in different semantic fields.[72]

Although the index is required for such cross-referencing in the print edition of *TOE*, electronic environments make it possible to search for senses located in other areas of the thesaurus with the click of a button.[73] In short, cross-references are, in various forms, present in historical language thesauri.

1.6.2. Editorial commentaries

Some of the thesauri covered in this chapter include editorial commentaries in their content proper. To illustrate, *DSSPIEL* provides an introduction per category, including bibliographical references that the reader can look into for further discussion of the lexical senses in the category. *LSM*, too, provides brief introductions and commentaries per category. Remarks per individual sense are provided in footnotes. Such introductions to, and commentaries on, specific sections of the thesaurus have been found most welcome.[74]

In contrast, scholars tend to receive an absence of editorial commentaries negatively. The rationale behind the structure or organisation of the thesaurus, for instance, is missing more often that not. As mentioned earlier, fashioning a

[70] Coleman, Review of *HTE1*, p. 209.
[71] Kay, Review of *ShT*, p. 74.
[72] Momma, Review of *TOE2*, p. 80.
[73] The current digital editions of *TOE* and *HTE* do not yet provide a means to quickly show the location of all polysemous senses of a categorized item without requiring the user to browse to the search engine and entering the word form there manually.
[74] Poultney, Review of *DSSPIEL*, pp. 331–2; Peters, Review of *LSM*, p. 399.

thesaurus involves subjective choices.[75] Although the result can be perceived in the form of the thesaurus itself, the rationale behind these choices is not always apparent. Why is this particular lexical sense found here and not elsewhere in the topical system? Why are co-ordinate categories presented in this specific order? Such questions have been posed by scholars reviewing historical language thesauri.[76] Unfortunately, the majority of these thesauri do not contain any such editorial remarks, or at least not on such a specific level, leaving it "to the reader's intellectual capacities or their creative guesswork to find explanations".[77] Finding editorial commentaries confined to an overall discussion as introduction to the thesaurus proper has left reviewers and users disappointed, desiring to acquire such illuminating information placed within the context of the very categories and entries of the thesaurus.[78] Since scholars encourage a higher inclusion of these constituents in the thesaurus proper, editorial remarks and commentaries are clearly valued in thesauri.

1.7. Related resources

Although few thesauri exist that cover a historical language, these lexicographic works have much in common with resources that are more commonplace. First, thesauri exist that do not cover a historical language but a contemporary one.[79] Second, semantic field studies, which are often smaller in scope than a thesaurus, not uncommonly adopt the same organizing principles as their larger counterparts.[80] Third, linguistic resources known as WordNets group sets of synonymous words and position these in formal, as opposed to informal, hierarchies.[81] That is to say, WordNets explicitly indicate hyponomy and meronymy in their semantic hierarchies rather than leaving the type of hierarchical relation between grouped words implicit. Fourth, indexing thesauri consist of one of the main components found in thesauri: the topical system.[82] These resources may be used to index content other than lexis, such as documents or audiovisual data. Lastly, dictionaries that arrange their lexis alphabetically contain many of the same components and their constituents as historical language thesauri do, be it with another organizing principle than a topical system.[83] Indeed, some thesauri have been fashioned through employing existing dictionaries as their source material.[84] As a result of these similarities between

[75] See section 1.3.

[76] Coleman, Review of *HTE1*, p. 209; Görlach, Review of *TOE1*, p. 398; Kay, Review of *ShT*, p. 73; Peters, Review of *LSM*, p. 400.

[77] Görlach, Review of *TOE1*, p. 400.

[78] Görlach, Review of *TOE1*, p. 400; Diller, Review of *HTE1*, p. 321; Coleman, Review of *HTE1*, p. 209; Peters, Review of *LSM*, p. 400.

[79] E.g., *Collins Thesaurus*.

[80] E.g., Diller, 'Emotions in the English Lexicon'.

[81] Fellbaum, 'WordNet and Wordnets'. In fact, the Open English WordNet is available as Linguistic Linked Data, too (see McCrae, 'English WordNet 2020').

[82] E.g., 'NASA Thesaurus' and 'Medical Subject Headings RDF'.

[83] E.g., *OED*.

[84] E.g., *TOE* has drawn from Clark Hall's *A Concise Anglo-Saxon Dictionary* and Bosworth and Toller's *An Anglo-Saxon Dictionary*; *HTE* from *OED*.

historical language thesauri and other kinds of resources, findings and conclusions on the dissemination of historical language thesauri provided by the dissertation may be relevant in a wider context.

1.8. Conclusion

This chapter has addressed what information components constitute the content of historical language thesauri. Through an analysis of eight such thesauri and of publications and handbooks on both thesauri and lexicography in general, three main components have emerged: (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. For each main component, a discussion of its presentation in thesaurus editions resulted in an overview of its constituents and ascertained what meaning, if any, can be attributed to that manner of their presentation. The order in which co-ordinate lexical senses are shown, for instance, has thus been found to convey no additional knowledge to the user. Instead, their co-ordination is based on information on the captured senses themselves (such as diachronic usage information), intended to offer users a familiar structure in which to find the knowledge they seek. These insights into the content of historical language thesauri are key in understanding their essence and determining how Web-based dissemination of these lexicographic works, and thesauri in general, can be improved so as to answer to the research needs of scholars in various disciplines.

References

Atkins, B. T. S. and M. Rundell, *The Oxford Guide to Practical Lexicography* (Oxford, 2008).

Bosworth, J. and T. N. Toller, *An Anglo-Saxon Dictionary Based on the Manuscript Collections of the Late Joseph Bosworth* (London, 1898), *Supplement* by T. N. Toller (Oxford, 1921), with *Enlarged Addenda and Corrigenda* by A. Campbell (Oxford, 1972).

Brewer, C., 'Authority and Personality in the *Oxford English Dictionary*', *Transactions of the Philological Society* 103.3 (2005), 261–301.

Brewer, C., Review of *HTE1*, *The Review of English Studies* 61.252 (2010), 801–5.

Brewer, C., 'Labelling and Metalanguage', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 488–500.

Collins English Thesaurus, eds. J. Crozier et al. (Glasgow, 2011).

BTH = The Bilingual Thesaurus of Everyday Life in Medieval England, eds. (Glasgow, 2019). https://thesaurus.ac.uk/bth/. Accessed: 15 March 2022.

Busse, B., 'A Celebration of Words and Ideas: The Stylistic Potential of the *Historical Thesaurus of the Oxford English Dictionary*', Review of *HTE1*

and *HTE2*, *Language and Literature* 21.1 (2012), 84–92.
Clark Hall, J. R., *A Concise Anglo-Saxon Dictionary*, 4th edn, with a supplement by H. D. Meritt (Cambridge, 1960).
Coleman, J., Review of *HTE1*, *Word Structure* 6.2 (2013), 201–13.
The Concise Scots Dictionary, ed. M. Robinson (Aberdeen, 1985).
Cruse, D. A., *Lexical Semantics* (Cambridge, 1986).
Dance, R., Review of *TOE1*, *Medium Ævum* 66.2 (1997), 312–13.
Diller, H., 'Emotions in the English Lexicon: A Historical Study of a Lexical Field', *English Historical Linguistics 1992: Papers from the 7th International Conference on English Historical Linguistics*, eds. F. Fernandez et al., Current Issues in Linguistic Theory 113 (Amsterdam, 1994), 219–34.
Diller, H., Review of *HTE1*, *Anglia* 128.2 (2010), 319–23.
DOE = *Dictionary of Old English: A to I Online*, eds. A. Cameron et al. (Toronto, 2018). https://tapor.library.utoronto.ca/doe/.
DOEC = *Dictionary of Old English Web Corpus*, eds. A. diPaolo Healey et al. (Toronto, 2009). https://tapor.library.utoronto.ca/doecorpus/.
DSSPIEL = Buck, C. D., *A Dictionary of Selected Synonyms of the Principal Indo-European Languages: A Contribution to the History of Ideas* (Chicago, 1949).
Faria, F., 'Georges Cuvier et le premier paradigme de la paléontologie', *Revue de Paléobiologie* 32.2 (2013), 297–302.
Fischer, A., 'The Notional Structure of Thesauruses', in *Categorization in the History of English*, eds. C. J. Kay and J. J. Smith, *Amsterdam Studies in the Theory and History of Linguistic Science* 4.261 (Amsterdam, 2004), pp. 41–58.
Fellbaum, C., 'WordNet and Wordnets', in: *Encyclopedia of Language and Linguistics*, 2nd edn, eds. Brown, K. et al. (Oxford, 2005), pp. 665-70.
Görlach, M., Review of *TOE1*, *Anglia* 116.3 (1998), 398–401.
Görlach, M., Review of *HTE1*, *English Language and Linguistics* 15.1 (2011), 193–7.
Hartmann, R. R. K., *Encyclopedia of Language and Linguistics*, 2nd edn, ed. K. Brown (Oxford, 2006), s.v. 'Thesauruses'.
Hartmann, R. R. K. and G. James, *Dictionary of Lexicography* (London, 1998).
Hausmann, F. J., 'Die Markierung im allgemeinen einsprachigen Wörterbuch: eine Übersicht', in *Wörterbücher: Ein internationales Handbuch zur Lexikographie*, eds. F. J. Hausmann et al. (Berlin, 1989), vol.1, pp. 649–57.
HTE = *Historical Thesaurus of English*; see *HTE1*, *HTE2*, *HTE3*.
HTE1 = *Historical Thesaurus of the Oxford English Dictionary*, 2 vols., eds. C. Kay et al. (Oxford, 2009).
HTE2 = *Historical Thesaurus of the Oxford English Dictionary*, eds. C. Kay et al. (2010). http://oed.com/thesaurus.
HTE3 = *The Historical Thesaurus of English*, eds. C. Kay et al. (Glasgow, 2014). http://historicalthesaurus.arts.gla.ac.uk.
HTS = *Historical Thesaurus of Scots*, ed. S. Rennie (Glasgow, 2015). http:

//scotsthesaurus.org.

Hüllen, W., *English Dictionaries, 800-1700: The Topical Tradition* (Oxford, 1999).

Hüllen, W., *A History of Roget's* Thesaurus: *Origins, Development, and Design* (Oxford, 2004).

Jacob, E. K., 'Classification and Categorization: A Difference that Makes a Difference', *Library Trends* 52.3 (2004), 515–40.

Kay, C., Review of *ShT*, *International Journal of Lexicography* 9.1 (1996), 71–6.

Kay, C., 'When Ignorance is Wisdom: Some Day-to-Day Problems of Classification', in *Categorization in the History of English*, eds. C. J. Kay and J. J. Smith, *Amsterdam Studies in the Theory and History of Linguistic Science* 4.261 (Amsterdam, 2004), pp. 59–69.

Kay, C. and M. Alexander, 'Diachronic and Synchronic Thesauruses', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 367–80.

Kirkpatrick, B., Review of *ScT*, *International Journal of Lexicography* 5.4 (1992), 305–9.

LSM = Coleman, J., *Love, Sex and Marriage: A Historical Thesaurus*, Costerus New Series 118 (Amsterdam, 1999).

McCrae, J. et al., 'English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology', Proceedings of the LREC 2020 Workshop on Multimodal Wordnets, Marseille, 11 May 2020, pp. 14–19. https://aclanthology.org/2020.mmw-1.3.pdf.

'Medical Subject Headings RDF', *U.S. National Library of Medicine*. https://id.nlm.nih.gov/mesh/. Accessed: 15 March 2022.

Momma, H., Review of *TOE2*, *Notes and Queries* 50.1 (2003), 79–80.

Murphy, M. L., *Encyclopedia of Language and Linguistics*, 2nd edn, ed. K. Brown (Oxford, 2006), s.v. 'Synonymy'.

Murphy, M. L., 'Meaning Relations in Dictionaries: Hyponymy, Meronymy, Synonymy, Antonymy, and Contrast', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 439–56.

'NASA Thesaurus', *Data.gov*. https://catalog.data.gov/dataset/nasa-thesaurus. Last updated: 12 November 2020.

Norri, J., 'Regional Labels in Some British and American Dictionaries', *International Journal of Lexicography* 9.1 (1996), 1–29.

OED = *Oxford English Dictionary*; see *OED1*, *OED2*, *OED3*, *OED Online*.

OED1 = *The Oxford English Dictionary*, eds. J. A. H. Murray et al. (Oxford, 1884–1933).

OED2 = *The Oxford English Dictionary*, eds. J. A. Simpson and E. S. C. Weiner, 2nd edn (Oxford, 1989).

OED3 = *The Oxford English Dictionary*, ed. J. A. Simpson. http://oed.com.

OED Online = *Oxford English Dictionary Online*. http://oed.com. Accessed: 18 October 2016.

Peters, H., Review of *LSM*, *Anglia* 120.3 (2003), 399–401.

Poultney, J. W., Review of *DSSPIEL*, *The American Journal of Philology* 71.3 (1950), 331–4.

Randell, D. A. et al., 'A Spatial Logic Based on Regions and Connection', Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning, Cambridge, Massachusetts, 25-9 October 1992, pp. 165–76.

Saeed, J. I., *Semantics*, 3rd edn, *Introducing Linguistics* 2 (Oxford, 2009).

ScT = *The Scots Thesaurus*, eds. I. Macleod et al. (Aberdeen, 1990).

Shakespeare, W., *Henry V*, ed. J. D. Mardock, Broadview Internet Shakespeare Editions (Peterborough, Ontario 2014).

ShT = Spevack, M., *A Shakespeare Thesaurus* (Hildesheim, 1993).

Stanojević, M., 'Cognitive Synonymy: A General Overview', *Linguistics and Literature* 7.2 (2009), 193–200.

Svensén, B., *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making* (Cambridge, 2009).

Taylor, P. P. G., *Encyclopedia of Language and Linguistics*, 2nd edn, ed. K. Brown (Oxford, 2006), s.v. 'Prototype Semantics'.

TOE = *A Thesaurus of Old English*; see *TOE1*, *TOE2*, *TOE3*, *TOE4*.

TOE1 = *A Thesaurus of Old English*, eds. J. Roberts et al., 2 vols., King's College London Medieval Studies XI (London, 1995).

TOE2 = *A Thesaurus of Old English*, eds. J. Roberts et al., 2 vols., Costerus New Series 131, 2nd edn (Amsterdam, 2000).

TOE3 = *A Thesaurus of Old English*, eds. J. Roberts et al. (Glasgow, 2005). http://libra.englang.arts.gla.ac.uk. Accessed: 1 February 2013.

TOE4 = *A Thesaurus of Old English*, eds. J. Roberts et al. (Glasgow, 2015). http://oldenglishthesaurus.arts.gla.ac.uk. Accessed: 15 March 2022.

Vrbinc, M. and A. Vrbinc, 'Diasystematic Information in the "Big Five": A Comparison of Print Dictionaries, CD-ROMS/DVD-ROMS and Online Dictionaries', *Lexikos* 25 (2015), 424–45.

W3C, 'HTML & CSS', *Web Design and Applications*. http://www.w3.org/standards/webdesign/htmlcss. Accessed: 3 August 2016.

Appendix 1.A:
Images of historical language thesauri

This appendix contains illustrative images of each of the historical language thesauri discussed in Chapter 1. Both an image of the thesaurus proper and its index is provided. For those thesauri that exist in multiple editions, only the latest paper edition and any electronic editions available online are illustrated here.

fore the world was'), and Aelfric gives it as the equivalent of Lat. *mundus*. OHG *mittilgart* and *weralt* are both freq. for *mundus* (e.g. Tat.; Otfr. only *worolt*).

ON *verǫld*, OE *weorold*, OHG *weralt*, etc. general Gmc. except Gothic, a cpd. of *wer* 'man' (2.21) and a noun form of adjs. for 'old' (14.15), hence 'age of man'. Thus it was equivalent to and followed Lat. *saeculum* in its ecclesiastical use (above, 2). For the early period this is the prevailing use, e.g. in the OE gospels, Mt. 12.32, 13.22 (Grk. αἰών, Vulgate *saeculum*). The further development was from this sense c to sense b and eventually to sense a, until the words in question became the common expressions for 'world' in all senses in all the Gmc. languages. NED s.v. *world*. Falk-Torp 1368.

5. Lith. *pasaulis*, Lett. *pasaule*, a cpd. of *pa-* 'under' and *saule* 'sun' (1.52), is 'world' in all senses. Lith. *pasaulis* is now preferred to the formerly current *svietas*, a Slavic loanword.

6. In Slavic there are two rival words for 'world', both still in common use in Russian, and both so used in dialects or older records of other Slavic languages.

ChSl. *mirŭ* (also *vĭsĭ mirŭ* with *vĭsĭ* 'all'), the regular translation of Grk. κόσμος, is the same word as *mirŭ* 'peace', used in Christian terminology as a collective 'community of peace' (cf. Russ. *mir*, also 'village community'), hence 'world' first in sense b. Berneker 2.60 f.

ChSl. *světŭ* 'light' (1.61) is used once in the gospels (Mt. 13.22) to translate Grk. ἀιών (above, 2). The semantic development was from 'light' to the 'realm of light, life' (cf. the association between 'light' and 'life', e.g. Grk. ὅστις φῶς ὁρᾷ 'whoever sees the light' = 'lives'), hence 'world' in sense c, later 'world' in all senses, as in the present Slavic languages.

7. Skt. *loka-* 'open space, earth, world' : Lat. *lūcus* 'grove', Lith. *laukas* 'field', Skt. *roca-* 'shining', Grk. λευκός 'bright', etc. Walde-P. 2.408 ff.

Skt. *jagat-* 'that which moves, lives, all creation, world', fr. *gam-* 'go'.

Skt. *bhū-*, *bhuvana-* 'existence, creation, earth, world', fr. *bhū-* 'be'.

Av. *gaēþā-* 'life, mankind, world', fr. the root of *gaya-* 'life', *jīva-*, Skt. *jīva-*, Lat. *vīvus*, etc. 'living' (4.74). Cf. Ir. *bith*, etc., above, 3.

Av. *aŋhu-* 'life, mankind, world', fr. *ah-* 'be' : Skt. *as-*, etc., IE **es-*.

Neither of these two Av. words, of which *gaēþā-* has the wider scope (Barth. 477 ff.), is actually quotable for 'world' in sense a, but only in senses b or c.

1.21 EARTH, LAND

Grk.	*γῆ, χθών*	Goth.	*airþa, land*	Lith.	*žeme*
NG	*γῆ*	ON	*jǫrð, land*	Lett.	*zeme*
Lat.	*terra, tellus, humus*	Dan.	*jord, land*	ChSl.	*zemlja*
It.	*terra*	Sw.	*jord, land*	SCr.	*zemlja*
Fr.	*terre*	OE	*eorþe, land*	Boh.	*země*
Sp.	*tierra*	ME	*erthe, land*	Pol.	*ziemia, ląd*
Rum.	*pamînt, ţară*	NE	*earth, land*	Russ.	*zemlja*
Ir.	*talan, tīr*	Du.	*aarde, land*	Skt.	*bhū-, bhūmi-, kṣam-*
NIr.	*talamh, tīr*	OHG	*erda, lant*		*pṛthivī-, mahī*, etc.
W.	*daear, tir*	MHG	*erde, lant*	Av.	*zam-, būmi*, OPers.
Br.	*douar*	NHG	*erde, land*		*būmi-*

Figure 1.A.1.: *DSSPIEL* thesaurus, p. 15.

INDEX OF HEADINGS[1]

[1] Also some words which, although not in the headings, denote related notions and are incidentally or partially covered in the discussion. In such cases the references are in parentheses.

Figure 1.A.2.: *DSSPIEL* index, p. 1505.

1. BIRDS, WILD ANIMALS, INVERTEBRATES

1.1 BIRDS

1.1.1 GENERAL

†aiten a partridge.
†awp a bullfinch.
bird, burd 1 a bird. 2 a young bird, nestling.
blackie the blackbird.
blethering Tam the whitethroat *WC, SW.*
blue bonnet the bluetit.
briskie, brichtie *now Kcb* the chaffinch.
bullie the bullfinch *now Fif.*
†buntlin the corn bunting.
bushock *Stlg,* **†bushsparrow** the hedge-sparrow.
capercailzie the wood-grouse.
cattle birds and beasts *in gen now Abd.*
chaffie a chaffinch *local.*
chitty wran the common wren *now Arg.*
†churr muffit the whitethroat.
chye the chaffinch *NE.*
†coal hood any of several species of black-headed birds.
corbie 1 the raven. 2 *occas* the carrion crow or hooded crow. 3 the rook *local.*
craw the crow, in Scotland *usu* applied to the rook.
†croupie the raven *chf Fif.*
craik the corncrake *now Abd Fif.*
cran, crane the swift *local C-S.*
croodlin doo a wood-pigeon *now Bnf Abd.*
croupie the common street pigeon *Fif Edb.*
cushat, cushie(-doo) *Gen except Sh Ork* the ring-dove, wood-pigeon.
cutty wran the wren *chf SW.*
deil, deevil, devil: ~'s bird the magpie *now Abd.*
doo a dove, pigeon.
feesant, fasiane, ephesian a pheasant.
feltieflier *now midLoth,* **feltifer** *now Fif,* **feltifare** *now Kcb* 1 *also* **feltie** *now midLoth Dmf* the field-fare. 2 the missel-thrush.
fern owl the nightjar.
†fewlume some kind of bird.
fieldy the hedge-sparrow.
gaiblick an unfledged bird *S.*
get an unfledged nestling *Gall Rox.*
goggie *child's word* an unfledged bird, a nestling *Ags SW.*
gorb *now Ags Kcb Uls,* **~el** *now NE Per Pbls,* **~lin** *now NE Ags,* **gordlin** *Abd,* **gorlin** *now Ags Per SW,* **~et** an unfledged bird. **~ellit** *of an egg* containing a developing chick *NE Per Pbls.*
gowd, goold, gold, golden crest(ie) the goldcrest *now midLoth.* **gowdie, goldie** *now nEC Lnk,* **gooldie** *now local Arg-S,* **~ spink** *now Per Pbls* the goldfinch.
gowk *now local NE-Uls,* **~oo** the cuckoo.
greenie, green lintie the greenfinch.
grund blackie a blackbird that nests on the ground *now midLoth.*
heather, hedder, hather: ~ blackie the ring ouzel.
hedge: ~ie, ~ spurdie, ~ spurgie a hedge-sparrow *now NE Per.*
hempie the hedge-sparrow *S.*
Hieland pyot the missel thrush *NE.*
huid, hood: ~ie craw, heidie craw *NE* 1 *also* **~ie, ~it craw, †~ock** the hooded crow. 2 *also* **~ie, ~it craw** the carrion crow.
jakie the jackdaw *EC Lnk SW.*
jay pyot a jay *now Kcb.*
kae the jackdaw.
Katie wren, kittie wren *now Loth Dmf Rox* the wren.
laverock, larick *now Mry Ags,* **†lairick** the sky-lark.
lintie, brown ~ *now Cai SW,* **gray ~** *now local* the linnet *Gen except Sh; see also* **rose lintie.**
lintwhite the linnet.
maggie the magpie *local Abd-S.*
mairtin martin, the housemartin *now Abd.*
†mak(e) a mate.
mason's ghost the robin redbreast *chf S.*
mavie mavis, the song-thrush.
†merl, *chf verse* the blackbird, *freq* **~ and mavis.**
mune, moon the goldcrest *Rox.*
nightingale, nichtingale the nightingale. **Scotch ~** the woodlark; the sedgewarbler *now C.*
†osill the ouzel, blackbird, merle; *perh also* the ring-ouzel.
ox ee 1 the great tit *now Fif midLoth.* 2 the bluetit *now nEC Rox.*
†paco(k) a peacock.
pai(r)trick a partridge *Gen except Sh Ork.*
†papejay the parrot.
papingo a parrot *latterly arch, hist, literary or as a heraldic device.*

Figure 1.A.3.: *ScT* thesaurus, p. 1.

INDEX

Figure 1.A.4.: *ScT* index, p. 477.

01.01
universe

creation, n.
monde, fr. n.
mundane, adj.
nature, n.
underworld, n.
universe, n.
world, n.
worldly, adj.
: 8

01.02
sky

caelum, lat. n.
celestial, adj.
ciel, fr. n.
cope, n.
element, n.
ensky, vb.
firmament, n.
heaven, n.
heavenly, adj.
horizon, n.
polus, lat. n.
roof, n.
sky, n.
skyey, adj.
skyish, adj.
welkin, n.
: 16

01.03
planet

aspect, n.
beneath world, n.
blazing star, n.
burning zone, n.
comet, n.
constellation, n.
crescent, n.
disaster, n.
disorb, vb.
Dragon's tail, n.
earth, n.
earthly, adj.
eclipse, n.
eclipse, vb.
equinoctial, n.
exhalation, n.
falling [star], n.
fixed star, n.
full moon, n.
globe, n.
globy, adj.
Guards, n.
half-moon, n.
heaven, n.
Hesperus
horn, n.
influence, n.
line, n.
lodestar, n.
lower world, n.
Mars
Mercury
meridian, n.
meteor, n.
middle earth, n.
moon, n.
moonish, adj.
new moon, n.
northern star, n.
north pole, n.
north star, n.
orb, n.
orbed continent, n.
planet, n.
planetary, adj.
pole, n.
polus, lat. n.
Saturn
seven stars, n.
shooting star, n.
sign, n.
Sol
sphere, n.
sphere, vb.
spherical, adj.
sphery, adj.
star, n.
star-like, adj.
starry, adj.
stelled, adj.
sun, n.
sunlike, adj.
sunny, adj.
tellus, lat. n.
terra, lat. n.
terre, fr. n.
terrene, adj.
terrestrial, adj.
under globe, n.
unsphere, vb.
unsunned, adj.
Venus
zenith, n.
zone, n.
: 74

01.04
zodiac

Aries
Bear
Bull
Cancer
Charles' wain, n.
Dragon
fiery Trigon, n.
Ram
Taurus
Ursa Major
Virgo
zodiac, n.
: 12

01.05
weather

climate, n.
climate, vb.
clime, n.
distemperature, n.
halcyons' days, n.
temperance, n.
temps, fr. n.
weather, n.
: 8

01.06
wind • storm

Aeolus
aer, lat. n.
aerial, adj.
air, n.
air, fr. n.
air, vb.
airless, adj.
airy, adj.
Aquilon
blast, n.
blastment, n.
blow, vb.
bluster, n.
bluster, vb.
blusterous, adj.
Boreas
drift-wind, n.
earthquake, n.
eastern wind, n.
east wind, n.
flaw, n.
four winds, n.
gale, n.
gust, n.
gusty, adj.
hurricano, n.
northeast wind, n.
north [wind], n.
overblow, vb.
sea-storm, n.
south, n.
south-west [wind], n.
south-wind, n.
storm, n.

Figure 1.A.5.: *ShT* thesaurus, p. 1.

A

Figure 1.A.6.: *ShT* index, p. 218.

A THESAURUS OF OLD ENGLISH

1. The Physical World

01 Earth, world: brytengrundas^{op}, brytenwangas^{op}, eormengrund^{op}, eorþe, eorþrīce, eorþstede^{op}, eorþweg^{p}, folde, foldwang^{p}, foldweg^{p}, grundwæg^{op}, grundwang^{p}, gumrīce^{p}, lēoht, middangeard, moldweg^{p}, woruld

.As God's creation: ealdgeweorc^{p}, frumhēowung^{og}, frumsceaft, frumgesceap^{op}, frumsetnung^{g}, frumweorc, fyrngesceap^{op}, fyrnweorc^{p}, landgesceaft^{op}, sǣ and eorþe, gesceaft, woruld, woruldgesceaft

..In the beginning: frymþ

..Awaiting redemption: gehæftworld^{o}

..Before the last judgement: ǣrworuld^{op}

.As dwelling place: eardgeard^{p}, ēþel, hrūse^{p}, molde, þēodeorþe^{o}, willwong^{op}, wundorworuld^{op}

.As distinct from heaven: grund, hrūse^{p}, woruld, woruldrīce

.Material world: woruld, ymbhwyrft

.Of the world: eorþcund, eorþcundlic, eorþlic, middangeardlic, woruldcund, woruldlic

.In earthly situation: hēr, in werþēode, on grunde, under sunnan, under wolcna hrōfe, under wolcnum

..(In)to this world/life: hider

.In earthly manner: eorþlīce

01.01 Surface of the earth: sǣ and eorþe, wang, wīdland^{p}

.Region: ende, eodor, grundscēat^{p}, scēat

.Zone, belt: gyrdel(s)

.Horizon: ēaggemearc^{op}

..Space enclosed by horizon: hring

01.01.01 Direction:

See 01.03.01.05.01 Winds; 05.10.05.04.09 Side, quarter, direction

01.01.01.01 North: norþdæl

.Northern part or place: norþdæl, norþende, norþhealf, norþhyrne, norþland, norþscēata^{o}, thila

.Northern: norþ, norþanweard, norþerne, norþeweard, norþheald, norþlang, norþlic^{og}

..More: norþerra

..Most: norþmest

Figure 1.A.7.: *TOE2* thesaurus, p. 1.

INDEX

The Index lists all Old English words and phrases in the Thesaurus, together with the lowest numbered headings under which they appear. In the alphabetical order of words *æ* follows *ae*, *þ* follows *t* and the *ge-* prefix is disregarded, as is *(ge)-*. Where variant forms are given, the sort otherwise ignores the presence of parentheses. Preference is given as a rule to normalized forms, but because variation (for example *i*, *ie* and *y*; or *æ* and *ea*) is embedded in the source dictionaries some spelling variation must be negotiated by the reader.

Figure 1.A.8.: *TOE2* index, p. 721.

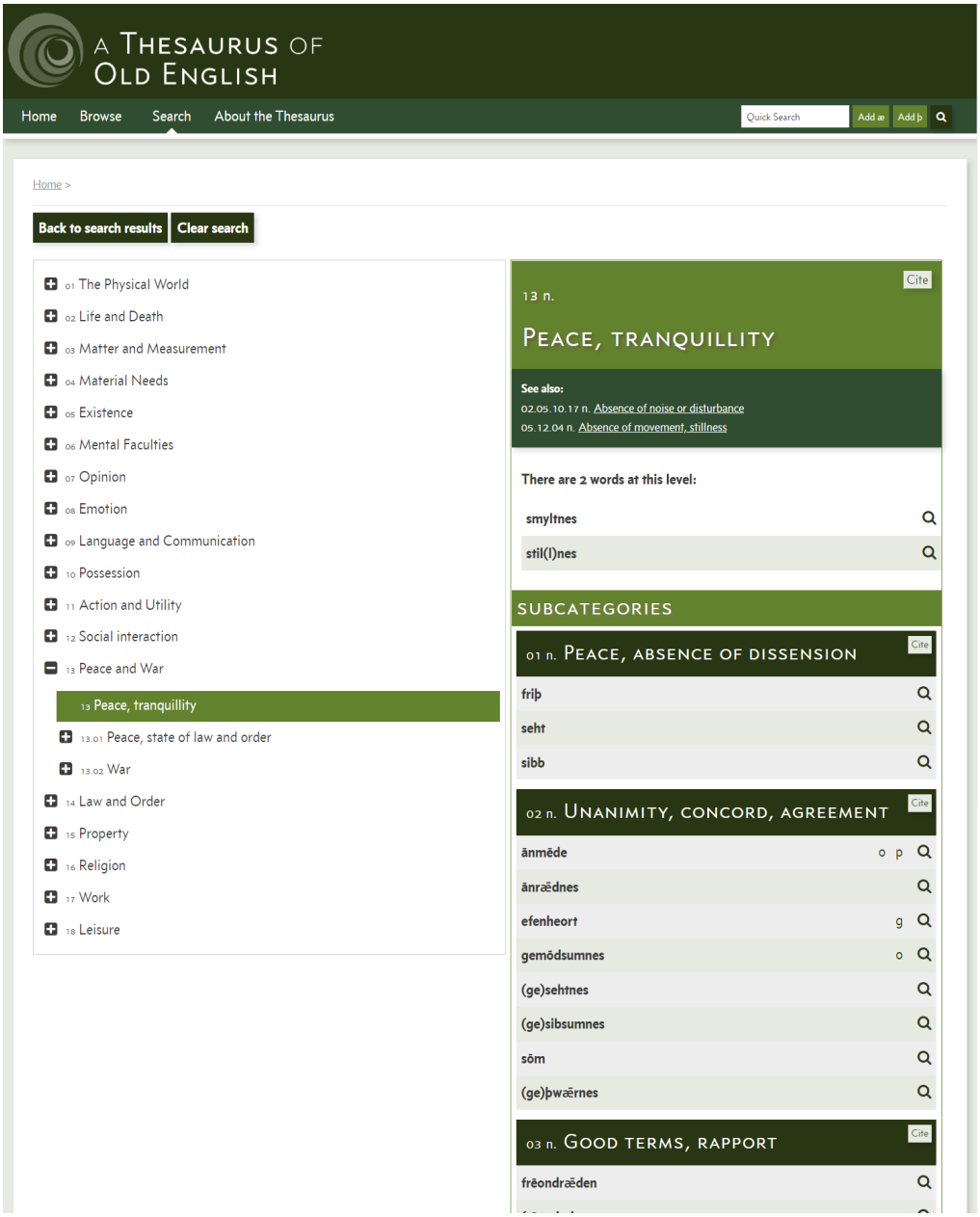


Figure 1.A.9.: *TOE4* thesaurus, category “13 n. Peace, tranquility”.

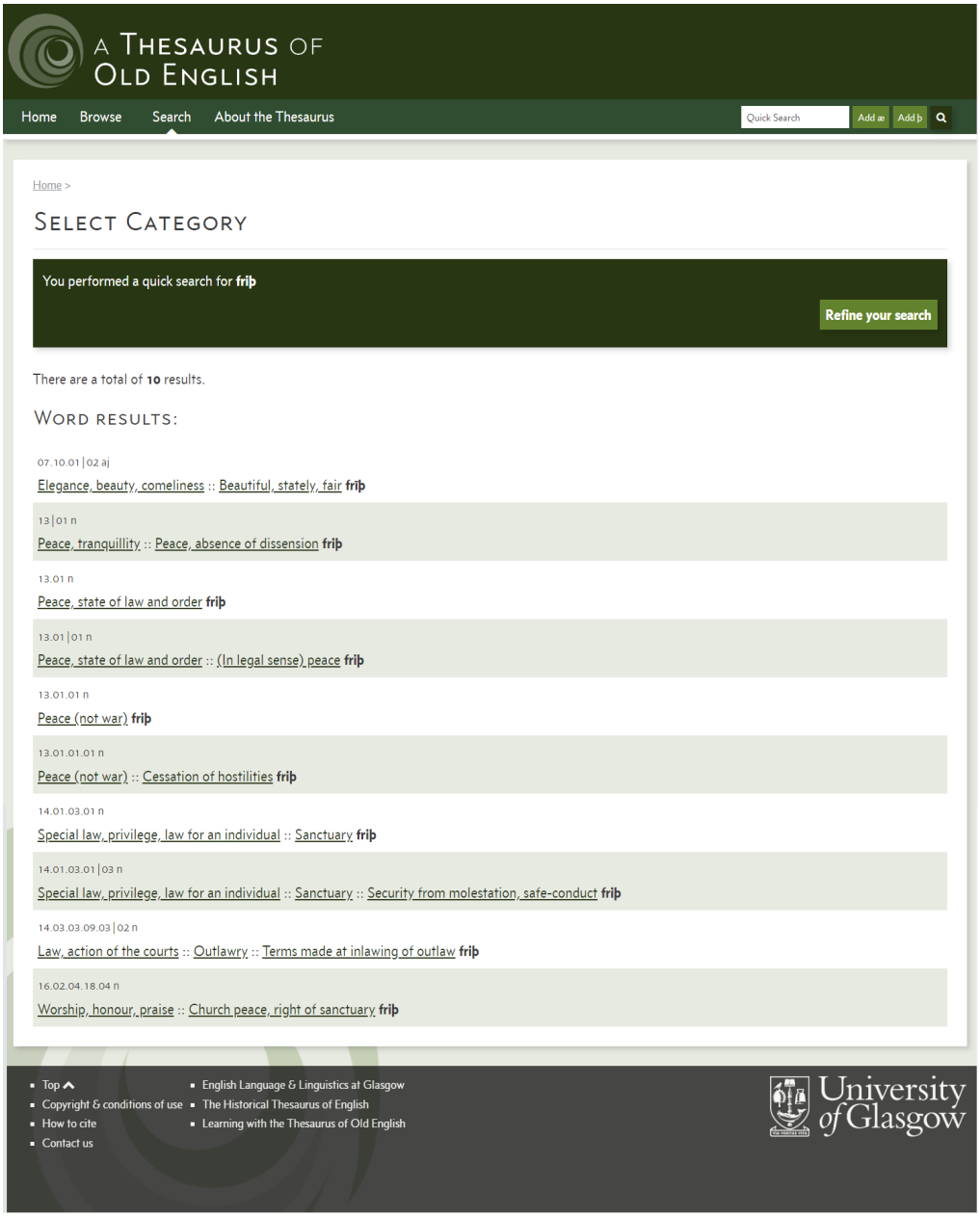


Figure 1.A.10.: *TOE4* index, 'friþ'.

LOVE

> Love itself, having lost its plain unsophisticate nature, and being now reduced into an art, has, like other arts, had recourse to particular words and expressions: of which it no more behoves lovers to be ignorant, than for seamen to be unacquainted with the terms of navigation. (*The Dictionary of Love* 1753: v)

L Love

see L/01 Types of Love, etc., L/02 Aspects of Love, etc., L.01 Loved One, etc. , L.02 Affection, L.03 Friend, L.04 Romantic Love, etc., M/07 Conjugal Love

> [love] defies the fickle smile of fortune, is neither warmed by summer's heat, nor chilled by winter's cold, but is itself unchangeable — true as life, and stronger than death. (A Lady *A Manual of the Etiquette of Love, Courtship and Marriage* 1853: 5-6)

† These terms could be defined more specifically, and some of them are in OE dictionaries, but there is no particular advantage, for example, in listing *heortlufu* separately as "hearty love" and *ferhðlufu* as "heartfelt love".

love. n: ferhðlufu† [OE nonce poet]; *freodscipe*[1] [OE nonce gloss]; freondlufu† [OE nonce poet]; freondræden OE; *freondscipe* [OE gloss]; heortlufu† [OE nonce poet]; *lufsumnes* [OE gloss]; modlufu† OE; myne OE; (ge)wilnung [OE derog]; lovered<lufræden OE - a1425 + [1456 - 1560 Sc]; love<lufu OE —; bosom c1200 - a1475; charity[2] a1225 - 1841; loveness a1240 + a1475; amour c1300 - 1742; loving 1372 - a1500; fair 1423; lust c1425 + a1450; well-liking (of) 1571 - 1679; beloving 1589; inloving[3] [1633 nonce]; affixedness [1668 derog]; fainness [1813 + 1827 + [1866 dic] Sc]; greeshoch [1821 + [1900 dic] Sc]; gra(h) [1848 - 1890 + [1900 dic] Ir]; luv [1898 — dial/colloq].
.of, like or pert. to love. adj: amative 1636 - a1678 + 1850 + 1971; amorevolous a1670; lovelike 1839 + 1992.
.to feel love. vb: deoran [OE poet]; (ge)freogan OE.

---

[1] This could be an alternative form of *freondscipe*, or based on *freod* "affection".
[2] Esp. associated with the results of love.
[3] Used to counter *injoying* (1633 W. Struther *True Happiness* 124).

Figure 1.A.11.: *LSM* thesaurus, p. 36.

Index

—A—

Figure 1.A.12.: *LSM* index, p. 448.

01 The world

01 (*n.*) *The world*
brytengrund OE · brytenwangas OE · eardgeard OE · eormengrund OE · eorþstede OE · eorþweg OE · eþel OE · foldwang OE · foldweg OE · grund OE · grundwæg OE · grundwang OE · gumrice OE · hruse OE · leoht OE · middan(g)eard OE · moldweg OE · þeodeorþe OE · wundorworuld OE · ymbhwyrft OE · earth-ric<eorþrice OE–*c*1200 · world-riche<woruldrice OE–1390 · fold<folde OE–*c*1440 · mould<molde OE–1614; 1810 (*poet.*) · earth<eorþe OE– · mother<modor OE– · world<woruld OE– · all the world *c*1175– · wone *c*1205–1748 · mound *a*1290–*c*1320 · wilderness *a*1340– · mappemonde *a*1390–1560 · the whole world 1534– · the globe of the world 1553 · the/this globe 1575– · mother earth *c*1586– · orb 1601–1667 · glimpses of the moon 1602; 1828– · the earthly globe *c*1630 · universe 1630–1820 · terrene 1667–1830 · a world 1676– · Orphic egg 1684–1789 · all outdoors 1830– (*US colloq.*) · outer world 1868– · Midgard 1882 · overworld 1911– · spaceship earth 1966– **01** *as God's creation* ealdgeweorc OE · frumgesceap OE · frumheowung OE · frumsceaft OE · frumsetnung OE · frumweorc OE · fyrngesceap OE · fyrnweorc OE · landgesceaft OE · sæ and eorþe OE · (ge)sceaft OE · woruld OE · woruldgesceaft OE · creature *a*1300–1611 · creation 1611–1783/94 **01.01** *before Christ's redemption* gehæftworld OE **02** *as capable of growth* vegetable *a*1676

01 (***adj.***) *Pertaining to world*
eorþcund OE · eorþcundlic OE · middangeardlic OE · woruldcund OE · woruldlic OE · earthly<eorþlic OE– · terrestre 1340–1663 · temporal *c*1375– · terrene *a*1400– · terrestrial 1432/50– · mundane 1475– · earthish 1536 · terrenal *a*1555–1588 · mundial *c*1560 (*Scots*); 1619 · nether 1590– · inferial 1591 · sublunary 1592– · earthy 1595–1869 · terreal 1598 · terrestrene 1599 · temporary 1603–1751 · mundal 1614–1631 · terranean 1653; *a*1900; 1939 · circumterraneous 1678 · earth-born 1702–1864 · terrean 1704 · terraneous *a*1711–1725 · terrigenal *a*1734 · worldly 1812 · telluric 1836– · tellurian 1846– · transglobal 1953– **01** *as formed of land and water* terraqueous 1658– · terraquean 1861

01 (***adv.***) *Earthly*
eorþlice OE · temporally *c*1380–*a*1716 · terrestrially 1604–1821 · mundanely 1824 **01** *situation* hider OE · in werþeode OE · under wolcna hrofe OE · under wolcnum OE · on (the) ground<on grunde OE–1742; 1883 · here<her OE– · under/beneath (the) sun<under sunnan OE– · an-erthe *a*1300; *c*1305 · on (the) mould *a*1310–1614 · here beneath 1340; 1576 · from/of hence 1382; 1611 · here below 1632; 1766 **02** *towards earth* an-erthe 1297; *c*1305

01.01 The earth

01.01 (*n.*) *The earth*
wang OE · widland OE · sea and earth<sæ and eorþe OE–*c*1366 · sea and land 1338– · sea and sand *a*1352–*a*1400 **01** *features of* geography 1737– · palæogeography 1881–

01.01 (***adj.***) *Pertaining to surface*
surficial 1892– **01** *originating above* epigenic 1882

01.01 (***adv.***) *In manner of surface*
surficially 1895–

01.01.01 (*n.*) *Region of the earth*
eodor OE · grundsceat OE · landscipe OE · sceat OE · woruldsceat OE · end<ende OE–*c*1450 · earth<eorþe OE–1628 · coast *a*1300–1667 · province *c*1330–1751 · plage *c*1386–1613 · nook *a*1400/50 · party *a*1400–1578 · part *c*1400– · region 1432/50– · climate 1477–1794 · corner 1535– · parcel 1582(2) · quart 1590 · climature 1604 · latitude 1632–1719 · district 1712– · latitudes 1760/2– · zone 1829– · area 1845– · natural region 1905– · sector 1943– **01** *containing characteristic elements* core area 1956–

01.01.01 (***adj.***) *Pertaining to region of the earth*
01 *belonging to/existing in particular region* local *a*1500–

01.01.01.01 (*n.*) *Hemisphere*
hemisphere 1551– · half-world 1605; 1866 **01** *eastern* Old World *a*1596; 1837– **01.01** *quality of* old-worldism 1887– · old-worldness 1895 · old-worldliness 1934 **02** *western* transmeridian 1500/20 (*Scots*) · new world 1555–

01.01.01.01 (***adj.***) *Pertaining to hemisphere*
hemispherical 1872– · hemispheric 1889– **01** *eastern* old-world 1877– · old-worldish 1886 **01.01** *northern region of eastern hemisphere* Palæarctic/Palearctic 1858– **02** *western* New World *a*1855– · neogean 1857

01.01.01.01 (***adv.***) *Towards hemisphere*
01 *western* newworldward 1935

01.01.01.02 (*n.*) *Quarter*
quarter *c*1391– · quart 1559

01.01.01.03 (*n.*) *Zone/belt*
gyrdel(s) OE · climate 1375 (*Scots*); *c*1391–1796 · line 1387–1837 · region 1551 · clime 1553–1697 · zone 1559– · belt 1808– **01** *character of* zonality 1909 **02** *denoting adjacency* sub- 1852– **03** *one of five* zone *a*1500– **03.01** *Arctic* Arctic 1569–1678 (*also fig.*) · frigid zone 1622– · snows 1844 **03.02** *subarctic* subarctics 1898 **03.03** *Antarctic* Antarctic 1366– **03.04** *tropical* hæte OE · torrid zone 1586– · burnt zone 1614 · burning zone 1713 · the tropics 1837– **03.04.01** *things associated with* tropicana 1960– **03.05** *subtropical* subtropics 1886– · semi-tropics 1908 **03.06** *climatic* clime 1542–1832 · zone 1599– · climate 1601– · homoclime 1916– (*Geography*) **03.06.01** *specific* temperate zone 1551– · horse latitudes 1777– · iceland 1842; 1889 · storm-area 1853– (*also fig.*) · cloud-belt 1860 · cloud-ring 1860–1862 · sun-land 1861 · snow-belt 1874– (*US*) · taiga 1888– (*Russian*) · storm-zone 1889 · storm-belt 1891 · cold pole 1909– **03.07** *time* time-zone 1892– **03.07.01** *boundary of* date-line 1880– · international date-line 1910– **03.08** *biogeographical* eard OE · range 1856– · realm 1876– **03.08.01** *sub-biogeographical* subregion 1864– · province 1877– · subzone 1888– **03.08.02** *specific* moose-yard 1800– · tsetse district 1851 · tsetse country 1853–1877 · Notogaea 1868– · floral zone 1870 · Ornithogaea 1874 · fly country 1892 · fly belt 1894 · fertile crescent 1914– · alterne 1920– · Wallacea 1928(2)– (*Zoogeographical*) · pollen-zone 1946; 1973 **03.08.03** *boundary* Wallace('s) line 1868– · continental divide 1869– · timberline 1896– · tree-limit 1953–

01.01.01.03 (***adj.***) *Pertaining to zone/belt*
co-climatory 1652 · zonic 1797 · zonal 1882– **01** *Arctic* Arctic *c*1391– · Zembl(i)an 1749–1806 · north-polar 1784 · arctogaeal 1870 **02** *subarctic* subarctic 1854– **03** *Antarctic* Antarctic 1366– · antarctical 1693 **04** *tropical* tropical 1698– · intertropical 1794– · tropic 1799–1875 · intratropical 1811– (*Geography*) · sub-equatorial 1909– (*also fig.*) **05** *bordering on tropics* subtropical 1865– **05.01** *characteristic of* semi-tropic(al) 1853– · subtropical 1863– · subtropic 1891– **06** *biogeographical* Hudsonian 1835– · palaeogaean 1857 · palaeotropical 1857 · Palaearctic 1857– · Nearctic 1858– · neotropical 1858– · neoarctic 1868 · subregional 1875 · neotropic 1877 · tropicopolitan 1878– (*Natural History*) · Notogaean 1879 · Indogaean 1885; 1899 · Notogaeic 1896 · Ornithogaean 1899 · ecogeographical 1956– · ecogeographic 1962– **07** *not geographically diverse* tautographical 1860

01.01.01.03 (***adv.***) *Towards/in manner of zone/belt*
zonally 1873– **01** *in manner of tropics* tropically 1852–

01.01.01.03 (***vt.***) *Make into zone/belt*
01 *make Arctic* arcticize 1853 **02** *make tropical* tropicalize 1885–1888

01.01.01.04 (*n.*) *Horizon*
eaggemearc OE · horizon *c*1374– · horizontal 1555(2) · weather-gleam/glim 1802– (*Scots & northern English*) · sea-horizon 1821– · verge 1822– (*poet.*) · sky-line 1824– · rim 1842– (*chiefly poet.*) · sea-line 1880– · sea-rim *a*1881– **01** *space enclosed by* hring OE · horizon 1559; 1771 **02** *area beyond horizon* underworld 1847–1890

01.01.01.04 (***adj.***) *Pertaining to horizon*
horizontal 1555– · horizontic 1651 **01** *part* quartered 1671 **02** *lacking horizon* horizonless *a*1839–1892 · unhorizoned 1888

01.01.02 (*n.*) *Geodetic references*
01 *latitude* parallel 1555– · height 1585–1694 (*Geography*) · latitude 1622– · degree 1647– · elevation 1686 **01.01** *of a place* latitude *c*1391– (*Geography*) **01.01.01** *in the past* palaeolatitude 1959– (*Geology*) **01.02** *equator* burning line 1553 · burnt line 1555 · equinox 1579–1728 (*Geography*) · equinoctial 1584–1813 · the line 1588– · equator 1612– · the girdle of the world 1626–1781 · palaeoequator 1960– (*Geology*) **01.02.01** *arc of* latitude of vertex *c*1850 **01.02.02** *magnetic* dip equator 1883 **01.02.03** *area of equatorial calm* the doldrums 1855– **01.03** *polar circle* polar circle 1551– · Antarctic circle 1556– · Arctic circle 1622– **01.04** *tropic* tropic 1527– · trope 1735 **01.04.01** *of Cancer* (tropic of) Cancer 1555– **01.04.02** *of Capricorn* (tropic of) Capricorn 1555– · hiemal line 1635 **01.05** *condition of being on the same latitude with* parallelism 1739 **02** *longitude* longitude *c*1391– · length 1581 **02.01** *in the past* palaeolongitude 1964– (*Geology*) **02.02** *as known to the ancients* longitude 1432/50 **02.03** *difference of longitude between places* longitude 1594 **02.04** *prime meridian* Greenwich meridian 1870– **03** *meridian* meridional line *c*1391–1834 · meridian *c*1391– · midday circle 1559 · meridian line 1559–1833 **03.01** *arc of* meridian of vertex *c*1850 **03.02** *magnetic* magnetic meridian 1704 (*Dict.*); 1832– **03.03** *in the past* palaeomeridian 1961– (*Geology*) **03.04** *state of being on/aligned with the meridian* meridionality 1664–1755; 1973 **04** *pole* Antarctic pole *c*1391– · Arctic pole *c*1391– · pole 1551– · magnes *a*1653 **04.01** *axis* hinge *a*1300–1697 **04.01.01** *movement* nutation 1715– · pendulation 1909– · polar wandering 1924– **04.02** *aurora* aurora pole 1881 **04.03** *magnetic* magnetic pole 1701– · palaeopole 1962– (*Geology*) **04.03.01** *line of* agonic line 1863– **05** *Antipodes* Antipodes 1549– · underworld 1868 **06** *radius* palaeoradius 1960– (*Geology*)

01.01.02 (***adj.***) *Geodetic references*
01 *latitude* latitudinal 1778– · latitudinous 1906– **01.01** *equator* equinoctial 1667; 1860 · equatorial 1713– **01.01.01** *on other side of equator* trans-equatorial 1900 **01.01.02** *pertaining to equatorial calms* calmy 1818 **01.02** *in the past* palaeolatitudinal 1964– (*Geology*) · palaeo-equatorial 1966– (*Geology*) **01.03** *on same line of latitude* parallel 1634(2) **01.04** *close to equator* (*of latitude*) low 1748– **01.05** *far from equator* (*of latitude*) high 1748– **02** *longitude* longitudinarian 1853 · longitudinal 1874 **03** *meridian* meridian *c*1391– · meridional 1555– **03.01** *crossing the meridian* transmeridional 1883–1892 **04** *pole* polar 1551– · polary 1559–1658 · circumpolar 1696– · polifugal 1740 · subpolar 1826 · poleward 1881– · bipolar 1896– **04.01** *crossing the pole* transpolar 1850– **04.02** *measured by magnetic north* (*of bearing*) magnetic 1744– (*Nautical & Aeronautics*) **05** *Antipodes* antipodal 1646– · Antipodean 1861–

Figure 1.A.13.: *HTE1* thesaurus, vol. 1, p. 5.

abiotic
adj. 01.02.00.09

abiotically
adv. 01.02.03

abiotrophic
adj. 01.02.01.01.04.17.01

abiotrophy
n. 01.02.01.01.04.17.01

Abipon(e)/Abipón
n. 01.02.07.07.50 | 03
n. 02.08.02.17 | 08.01
adj. 01.02.07.07.50 | 02
adj. 02.08.02.17 | 15.01

Abiponian
n. 01.02.07.07.50 | 03
adj. 01.02.07.07.50 | 02
adj. 02.08.02.17 | 15.01

abite
vt. 01.02.08.01.27

Abitur
n. 03.06.05.03 | 06.01
adj. 03.06.05.03 | 04

abiturient
n. 03.06.03.02.02 | 06

abject
n. 03.01.04.06.04 | 08.01

abjection
n. 02.02.29.02 | 05
n. 02.06 | 04
n. 03.01.04.06.04 | 08.01

abjective
adj. 03.05.05.01.01

abjectly
adv. 02.02.29.05

abjunction
n. 01.02.03.05.06 | 04

abjurer
n. 02.06 | 04.04

abjure the realm/town
vi. 02.08.03.13.02 | 10

Abkhasian
adj. 01.01.06.02 | 04

Abkhasian/Abkhazian
n. 01.02.07.07.19

Abkhaz/Abkhas
n. 01.02.07.07.19
n. 02.08.02.05 | 02
adj. 01.02.07.07.19

ablactation
n. 01.02.08.03.05.03.01 | 07.04

ablantive
n. 02.08.04.01.04 | 14

ablaqueate
vt. 01.02.08.03.07 | 05

ablaqueation
n. 01.02.08.03.07 | 14

ablastemic
adj. 01.05.02.01.01

ablate
vi. 01.05.07.02.03 | 02.01
vi. 01.05.07.04.06 | 02.04
vt. 01.01.10.04.01 | 05
vt. 01.02.01.03.04.20.01
vt. 01.05.07.04.06 | 10
vt. 03.09.05.10 | 02.01

ablated
adj. 01.01.10.04.01 | 06
adj. 01.05.07.04.06 | 02

ablation
n. 01.01.05.07.02.01 | 24
n. 01.01.07.02.06 | 10
n. 01.01.10.04.01 | 15
n. 01.02.01.03.01.01 | 09
n. 01.02.01.03.04.20.01
n. 01.05.07.04.06 | 00; 07
n. 03.09.05.10 | 05.04.02

ablatitious force
n. 01.01.10.05.03.05 | 10

ablatival
adj. 02.08.04.01.04 | 10

ablative
adj. 01.02.01.03.04.20.01
adj. 01.05.07.04.06 | 02.02

ablator
n. 03.09.05.12.07.02 | 06.05

ablaut
n. 02.08.04.03.01.02 | 21

(ablaut-)series
n. 02.08.04.03.01.02 | 21.02

ablaze
adj. 02.02.15.01 | 04
adj. 02.02.17
adv. 01.04.03.03.02.03.02
adv. 01.04.08.01 | 02

able
n. 03.08.07.02.03 | 02
adj. 01.05.05.22 | 00; 10

able-bodiedness
n. 01.02.01 | 04

abled
n. 01.02.01 | 04.01
adj. 01.02.01 | 03.01
adj. 01.02.04.06.01 | 01
adj. 01.05.05.22

ablegate
n. 03.07.02.05.01.01.03 | 06

able(-)ism
n. 03.01.05 | 02.07

ableist
adj. 03.01.05 | 02.07

ablemost
adj. 01.05.05.22 | 01

ableness
n. 01.05.05.22

be able to use
vt. 02.07.04 | 01.01

ablette
n. 01.02.06.16.07.03 | 05.05.05.02

abloom
adj. 01.02.04.10.14.01.02 | 01.04

ablow
adj. 01.01.11.02.07 | 04
adj. 01.02.04.10.14.01.02 | 01

abluent
adj. 01.02.11.01.01 | 01

ablush
adj. 01.04.09.07.03

ablute
vi. 01.02.11.01.01.01
vt. 01.02.11.01.01.01

abluted
adj. 01.02.11.01.01 | 03

ablution
n. 01.02.11.01.01.01
n. 01.02.11.01.01.05 | 01
n. 01.04.02.13 | 09.17
n. 03.07.04.15.07.02 | 01

ablutionary
adj. 01.02.11.01.01.01

ablution bench
n. 01.02.11.01.01.01 | 02.03

ablution block
n. 01.02.11.01.01.01 | 02.01

ablution cabinet
n. 01.02.11.01.01.01 | 02.03

ablution place
n. 01.02.11.01.01.01 | 02

ablutions
n. 01.02.11.01.01.01 | 02.01

ablution trough
n. 01.02.11.01.01.01 | 03.02

ably
adv. 01.05.05.22

A.B.M.
n. 03.03.16.01.06.04 | 01

abmigration
n. 01.02.06.19.03 | 13

abnegator
n. 02.05.05.05.02 | 06.02

Abney clinometer
n. 01.01.09.01.02.02 | 17.10

Abney level
n. 01.01.09.01.02.02 | 17.10

abnormal
adj. 01.05.05.21.01 | 05
adj. 01.06.03.02.02.01.01

abnormalism
n. 01.06.03.02.02.01.01

abnormality
n. 01.06.03.02.02.01.01 | 00; 01

abnormalize
vt. 01.06.03.02.02.01.01

abnormally
adv. 01.06.03.02.02.01.01

abnormal psychology
n. 01.02.01.03.02 | 35

abnormously
adv. 01.06.03.02.02.01.01

aboard
adv. 03.09.01.06.02
adv. 03.09.03.03 | 01
adv. 03.09.04 | 03; 04
adv. 03.09.04.13.17.01.04 | 01
adv. 03.09.05.03 | 01
prep. 01.05.07.05.12 | 02
prep. 03.09.01.06.02
prep. 03.09.03.03 | 01
prep. 03.09.04 | 01
prep. 03.09.05.03 | 01

abode
n. 03.02
n. 03.02.07.02
n. 03.02.07.02.06
n. 03.11.03.03.05.11.02 | 31
vi. 02.01.14.09.05.06
vt. 02.01.14.09.05.06

abode of love
n. 03.02.07.02.05 | 06

abohm
n. 01.04.07.03.06.10 | 03

aboideau/aboiteau
n. 01.01.05.06.01 | 14.03.05

aboil
adv. 01.04.03.03.02.02 | 01

aboiteau
n. 01.01.05.06.01 | 14.03.05

abolishable
adj. 02.01.11.05.02 | 03

abolisher
n. 01.05.02.02 | 02

abolition
n. 02.01.11.04.01
n. 03.04.10.03.03 | 06.03

abolitionary
adj. 02.01.11.05.02

abolitionism
n. 03.04.10.03.03 | 06.03.01

abolitionist
n. 01.05.02.02 | 02
n. 03.01.05.05 | 08
n. 03.04.06.17.04.08 | 05.01
n. 03.04.10.03.03 | 06.03.02
n. 03.04.12.03 | 03.01

abolitionize
vt. 03.04.10.03.03 | 06.01

abolla
n. 01.02.10.02.06 | 21.01.02.01

abomasum/abomasus
n. 01.02.06.20.06.05 | 01.01.03.04

abominable snowman
n. 01.07.03.04 | 15.04

abominably
adv. 02.02.23.04

abomination
n. 02.02.23.04

abominationly
adv. 02.02.23.04

abominator
n. 02.02.23.03 | 03

abonné
n. 03.10.13.18.02 | 10.02

abonnement
n. 03.04.10.04 | 18.02.02
n. 03.10.13.18.02 | 10.03
adj. 03.04.10.04 | 04
adj. 03.10.13.18.02 | 02.03

aboral
adj. 01.02.06.12.02 | 02

aborally
adv. 01.02.06.12.02 | 03

abord
n. 01.05.08.05.12 | 09
vt. 01.05.08.05.12 | 07

abordage
n. 03.03.17.02 | 02

aborigen/aborigin
n. 03.02.06.02 | 01

aboriginal
n. 03.02.06.02 | 01
adj. 03.02.06.02 | 01; 01.01

Aboriginal/aboriginal
n. 01.02.07.07.46.03
adj. 01.02.07.07.46.03

aboriginalism
n. 03.01.05.05 | 06

aboriginality
n. 03.02.06.02 | 01.01

aboriginally
adv. 01.06.03.03.02 | 01

aborigine
n. 01.02.07.07.46.03
n. 01.02.07.08.03 | 05

ab origine
phr. 01.06.03.03.02 | 01

aborigines
n. 01.02.04.04 | 03
n. 01.02.06.02 | 01.01
n. 03.02.06.02 | 01

aborning
adj. 01.02.00.05 | 01
adv. 01.02.00.05

aborsement
n. 01.02.01.01.04.18.17 | 08

abort
n. 01.02.02.01 | 02.01
vi. 01.02.01.03.04.20.09 | 01.01
vi. 01.02.03.03.15 | 01.03
vt. 01.02.01.03.04.20.09 | 13.01; 13.01.01; 13.01.02
vt. 01.02.03.03.16.09 | 00; 01
vt. 01.05.05.19 | 07
v.refl. 01.02.03.03.16.09

aborted
adj. 01.05.05.19 | 08

abortee
n. 01.02.03.03.16.09 | 02

abortifacient
adj. 01.02.00.05.01 | 02.07

abortion
n. 01.02.01.01.04.18.17 | 08
n. 01.02.03.03.16.09

abortive
n. 01.02.01.01.04.18.17 | 08
n. 01.02.01.03.05.03.26 | 02
n. 01.02.02.01 | 02.01
n. 01.02.06.10 | 04.06
n. 03.08.07.08.01.02 | 01.04
adj. 01.02.00.05.01 | 02.07
adj. 01.02.01.01.04.18.17 | 03; 04
adj. 01.02.03.03.15 | 01.04
adj. 01.02.03.03.16.01
adj. 01.05.05.19
vt. 01.05.02.01.01

abortively
adv. 01.02.01.01.04.18.17 | 01
adv. 01.05.05.17.03.01.03 | 06
adv. 01.05.06.09.02 | 02

abortiveness
n. 01.05.05.19

abortment
n. 01.02.01.01.04.18.17 | 08

abortorium
n. 01.02.01.03.09 | 01.19.01

abortus
n. 01.02.03.03.16.09 | 03

abortus fever
n. 01.02.01.01.05.03 | 12

à bouche
adj. 03.08.04.07.01.01 | 02

aboulia
n. 01.02.01.02.01.02 | 28.03

aboulomania
n. 01.02.01.02.01.02 | 28.03

abound
vi. 01.06.06.04.01
vt. 01.06.06.04.01 | 06

abounder
n. 01.06.06.04.01 | 07.01
n. 02.07.05 | 11

abounding
adj. 01.06.06.04.01

abound with/in
vt. 01.06.06.04.01 | 06

about
adv. 01.05.05 | 01
adv. 01.05.06.08.05.02
adv. 01.05.06.11.03.01
adv. 01.05.07.03.06.03 | 01.01
adv. 01.05.07.06 | 11
adv. 01.05.08 | 01
adv. 01.05.08.05.02 | 01.01
adv. 01.06.06.02
adv. 03.09.04.03.03 | 06
vt. 03.09.04.03.03 | 11
prep. 01.05.03.01 | 02
prep. 01.05.07.01.02 | 04; 05
prep. 01.05.07.05.03.03
prep. 01.05.07.05.25.05 | 00; 01
prep. 01.05.08.02.05 | 01
prep. 01.05.08.05 | 01
prep. 01.06.06.02
prep. 02.01.12.08.02.01.01
prep. 03.01.04.01

be (all) about
vt. 02.01.07.04.03 | 01
vt. 02.01.10.02 | 02

be on about
vt. 02.08.03.04 | 03

about face
n. 01.05.07.06 | 18.02
n. 03.03.07.02 | 04
vi. 01.05.07.06 | 07.03
vi. 03.03.07.02 | 04
int. 03.03.09.04 | 08

about-forth
adv. 01.05.07.05.03.03

about one
phr. 02.01.08.01.02 | 01

about-sledge
n. 03.10.11.07 | 03.04

about-speech
n. 03.11.03.03.05.04 | 10.01

about time
adv. 01.05.06.09.02 | 03

about to
adv. 01.05.01.07.02 | 02

about turn
int. 03.03.09.04 | 08

about-turn
n. 01.05.07.06 | 18.02
n. 02.01.13.03.04
vi. 01.05.07.06 | 07.03

aboutward
adv. 01.05.01.07.02 | 02
adv. 01.05.05.10 | 01

about-writing
n. 03.10.13.15.05.01 | 22.02.01

above
adj. 01.05.06.08.04.03 | 14
adj. 02.08.03 | 04.02
adv. 01.01.04.04.01
adv. 01.01.05.01.01 | 01
adv. 01.04.03.03.01 | 01
adv. 01.05.05.16.02.01
adv. 01.05.07.05.03.01
adv. 01.05.07.05.07 | 00; 01
adv. 01.06.06.03 | 02.01
adv. 01.06.06.05.01 | 02
adv. 01.07.04.06 | 01
adv. 03.01.06.01.02 | 01
adv. 03.11.03.03.01.01 | 02
prep. 01.05.07.02.04.06 | 01
prep. 01.05.07.05.07 | 01
prep. 01.06.06.03 | 01

at/to one's above
phr. 01.05.05.16 | 05

above and beyond
phr. 01.06.06.05.01

above/beyond the nock
phr. 01.06.06.04.02 | 04

above board
adv. 02.01.12.08.04
adv. 03.08.02.01.01

above-cited
adj. 02.08.03 | 04.02

above-found
n. 01.06.04.08 | 01

above-given
n. 01.06.04.08 | 01

above ground
adv. 01.01.04.01 | 01
adv. 01.02.00 | 01
adv. 03.11.05.08 | 01

above-hand
adv. 01.05.08.09.01 | 01

above-head
adv. 01.05.07.05.07 | 01

above measure
phr. 01.06.06.04.02 | 04

above-mentioned
adj. 02.08.03 | 04.02

above-named
adj. 02.08.03 | 04.02

above one's bend
phr. 01.05.05.11.02 | 01

above oneself
adv. 01.02.06.20.06.02.07 | 02
adv. 02.02.17 | 01

above one's match
adj. 02.01.15.07.08.04

above one's reach
phr. 01.05.05.11.02 | 01

above one's thumb
phr. 01.05.05.11.02 | 01

above/over the weather
adv. 03.09.05.01.01 | 02

above par
adj. 02.01.15.07.08.04 | 03

abovesaid
adj. 02.08.03 | 04.02

above stairs
adv. 03.02.07.03.06 | 02

above suspicion
adj. 02.01.13.02.02 | 02
adj. 02.01.15.07.08 | 11

above the common
phr. 01.06.02.02.04 | 04

above (the head of)
adv. 02.01.09 | 01

above the line
adj. 03.11.01.05.02.07.02 | 08

above the moon
adv. 01.01.10.01 | 01
phr. 01.06.06.04.02 | 04

above the rest
phr. 01.06.02.02.04 | 02

above the salt
adv. 03.01.06.01.01

above water
phr. 01.04.05.11 | 02

above-written
adj. 02.08.03 | 04.02

ab ovo
phr. 01.06.03.03.02 | 01

abox/a-box
adv. 03.09.04.13.18.01.01 | 01

abracadabra
n. 01.07.02.02.03 | 02.03.01

abrade
vi. 02.01.11.05

abrader
n. 01.05.02.02.06 | 07

Abraham/Abram man
n. 03.09.00.05 | 04.09.03

Abrahamic
adj. 03.01.01.03.10 | 01.01
adj. 03.07.00.18.03.01

Abraham's Balm
n. 01.02.04.13.02.03.22

in Abraham's bosom
adv. 01.07.04.06 | 01

abraid
n. 02.01.18.06.04 | 04
vi. 01.02.01.01.04.18.03 | 03.02
vi. 01.03.01.02 | 04.02
vi. 01.05.08 | 02
vi. 03.02.02.01
vt. 02.01.18.06.04

abranchial
adj. 01.02.06.12.02.06 | 06.06

abranchiate
adj. 01.02.06.12.02.06 | 06.06

abrasional
adj. 01.01.07.02.06

abrasion platform
n. 01.01.04.03.05.01 | 02

abrasive
n. 01.05.02.02.06 | 07
adj. 03.11.03.03.05.10 | 09

Figure 1.A.14.: *HTE1* index, vol. 2, p. 2.

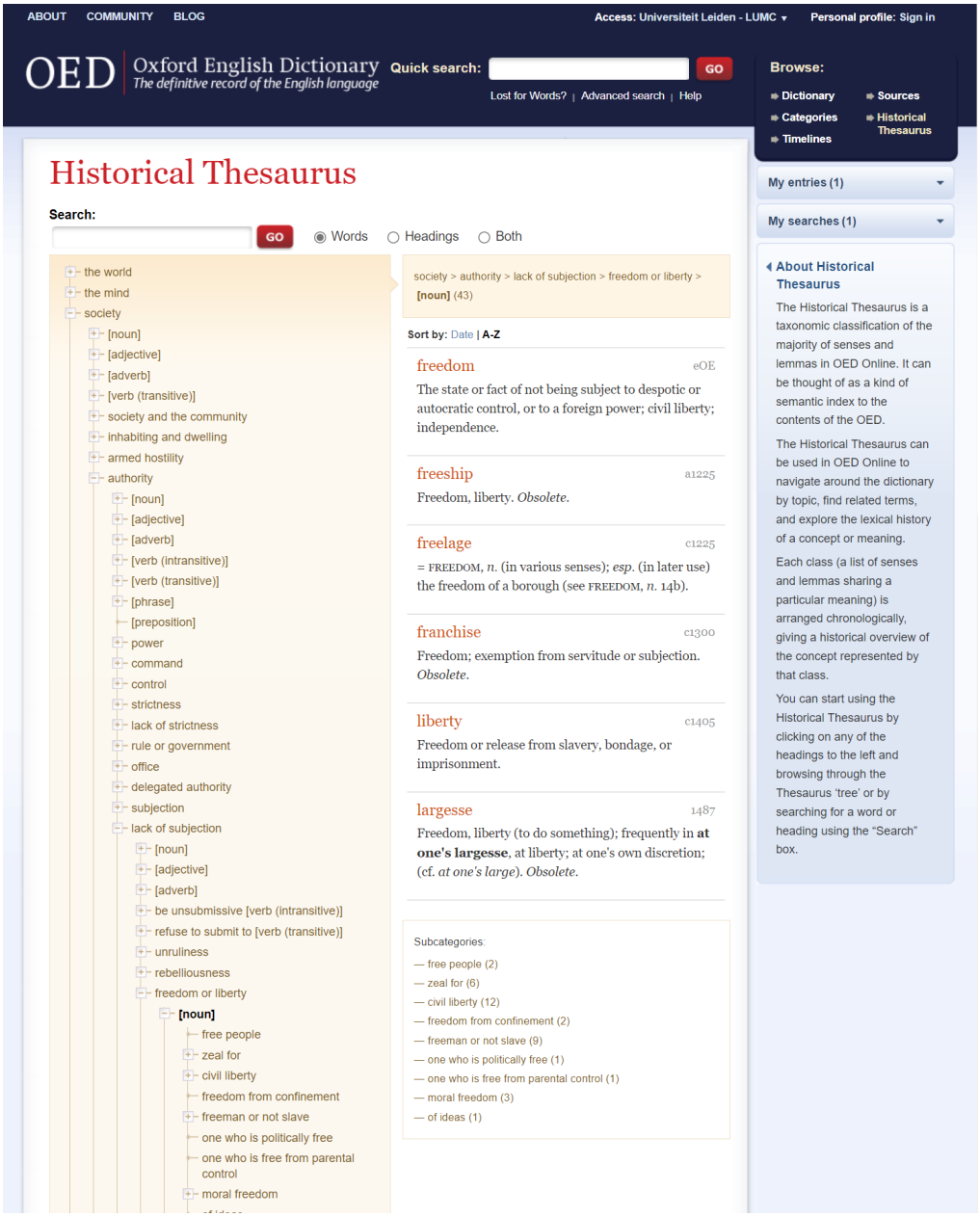


Figure 1.A.15.: *HTE2* thesaurus, category "freedom or liberty".

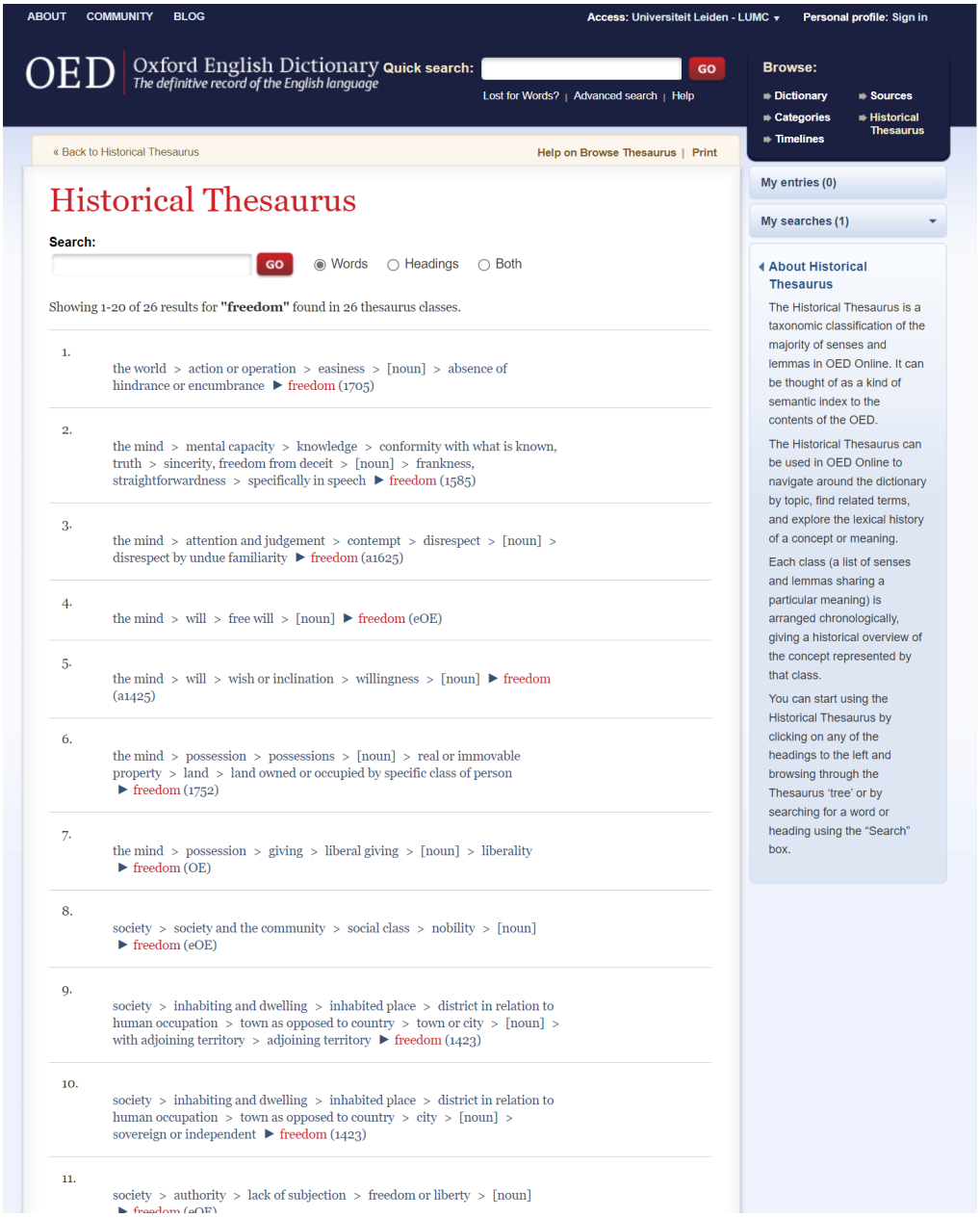


Figure 1.A.16.: *HTE2* index, 'freedom'.

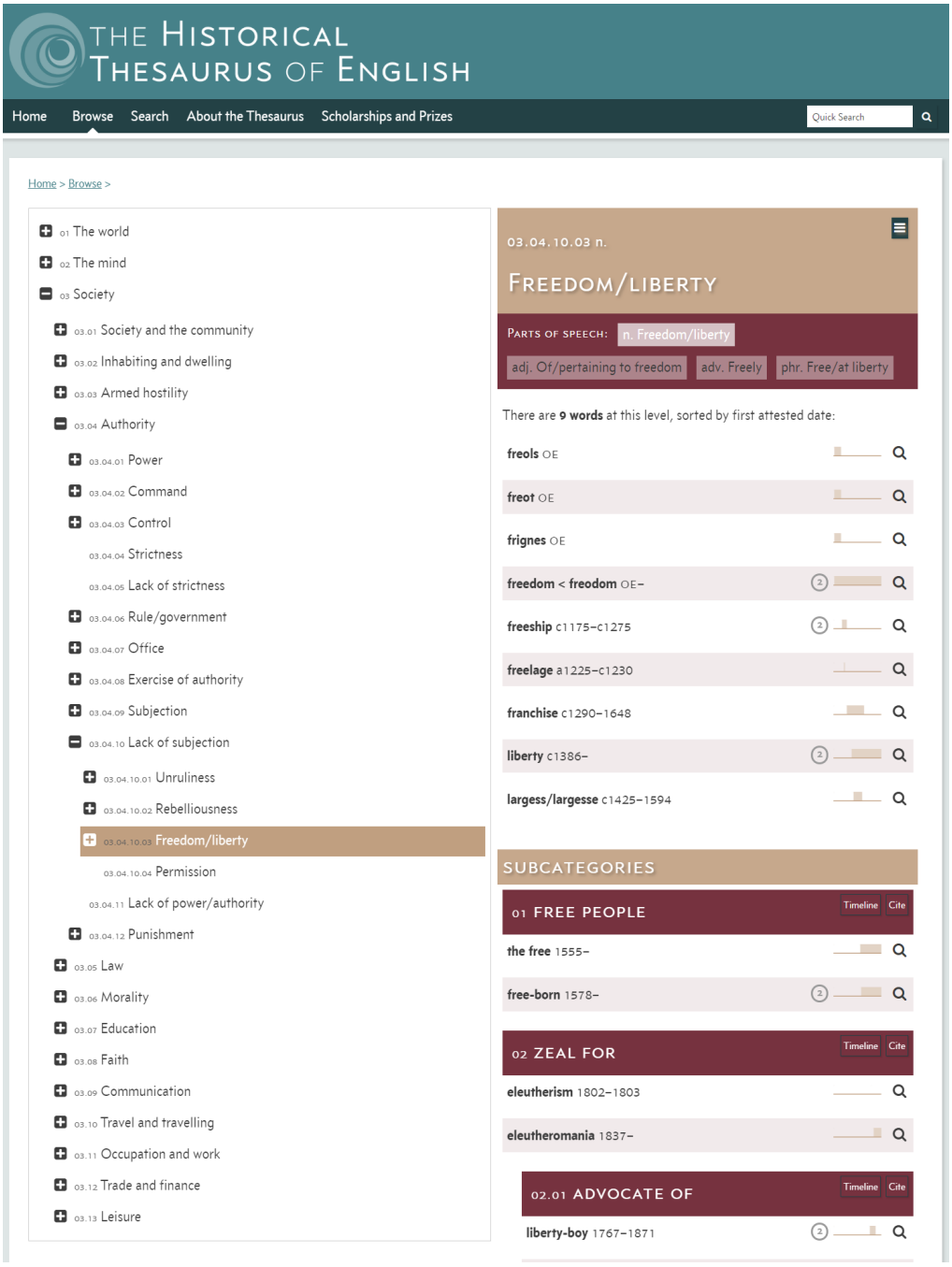


Figure 1.A.17.: *HTE3* thesaurus, category "03.04.10.03 n. Freedom/liberty".

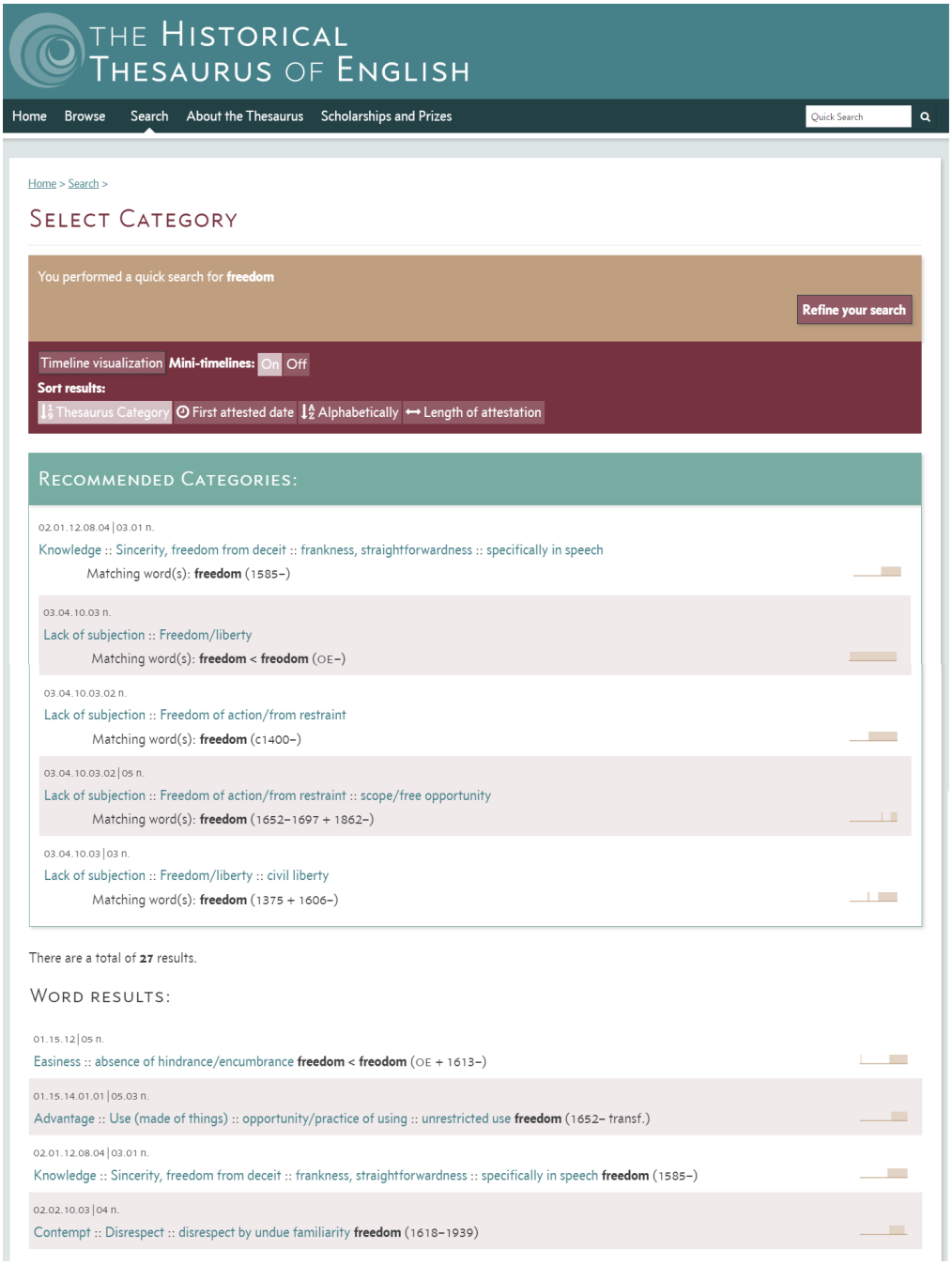


Figure 1.A.18.: *HTE3* index, ‘freedom’.

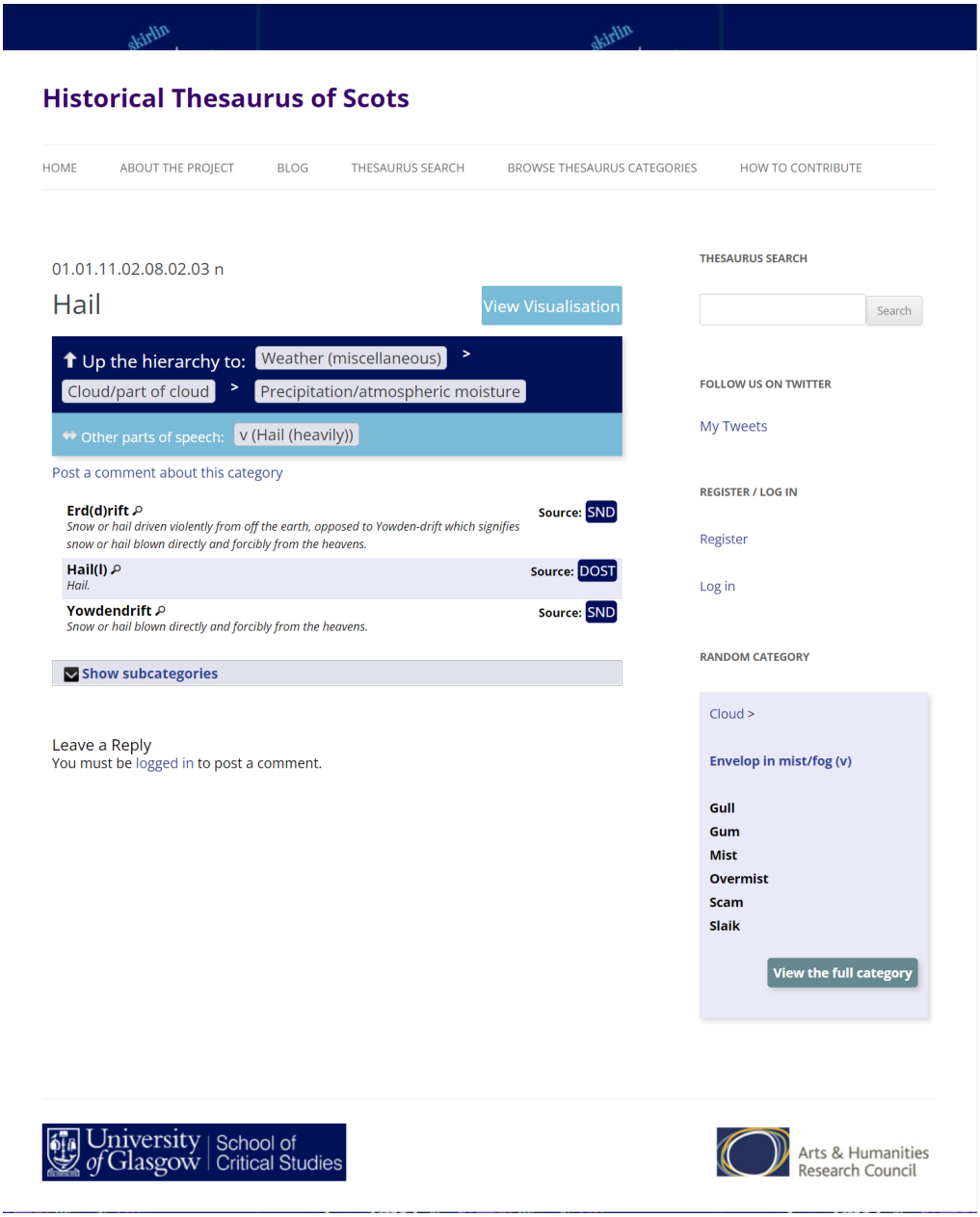


Figure 1.A.19.: *HTS* thesaurus, category "01.01.11.02.08.02.03 n. Hail".

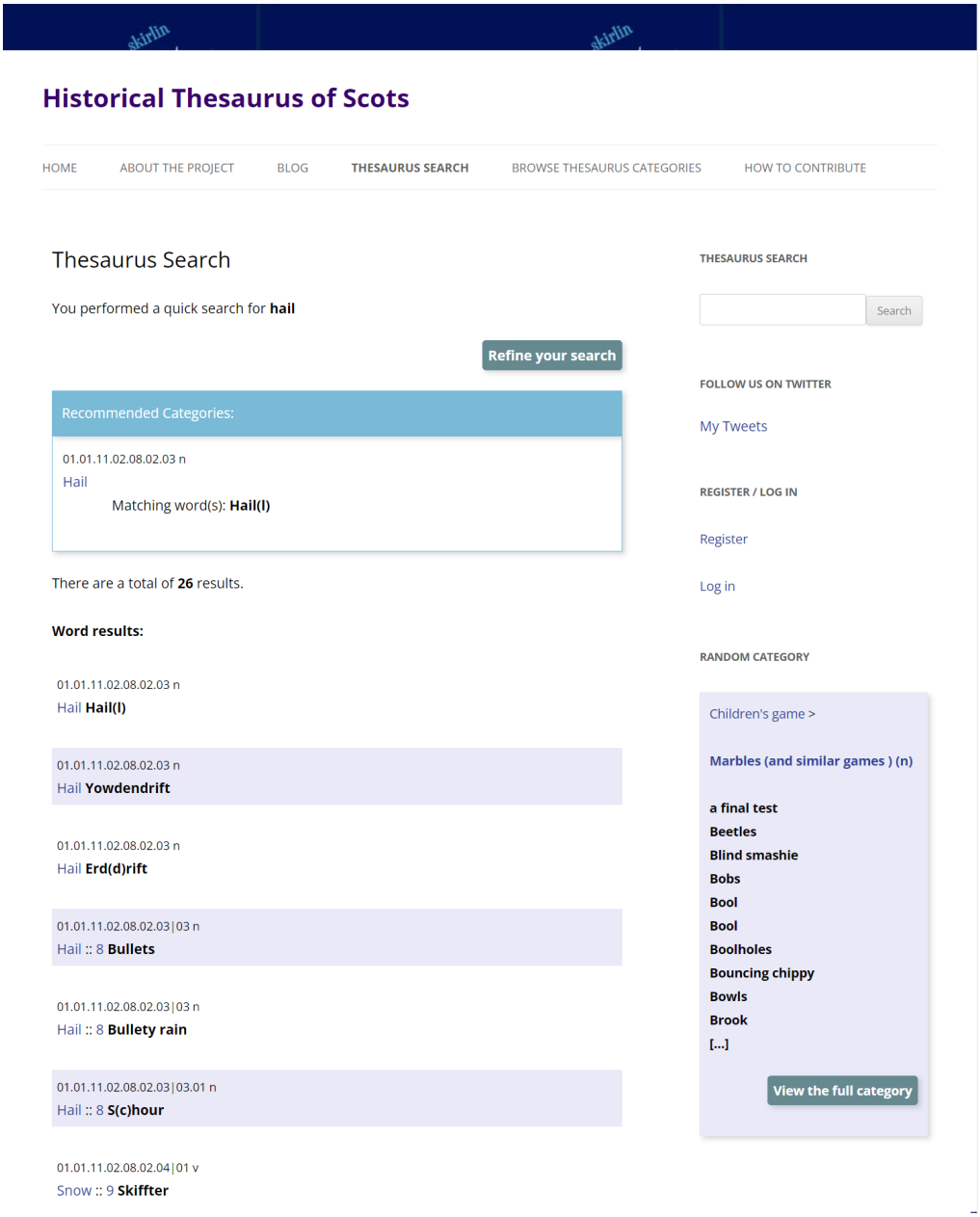


Figure 1.A.20.: *HTS* index, 'hail'.

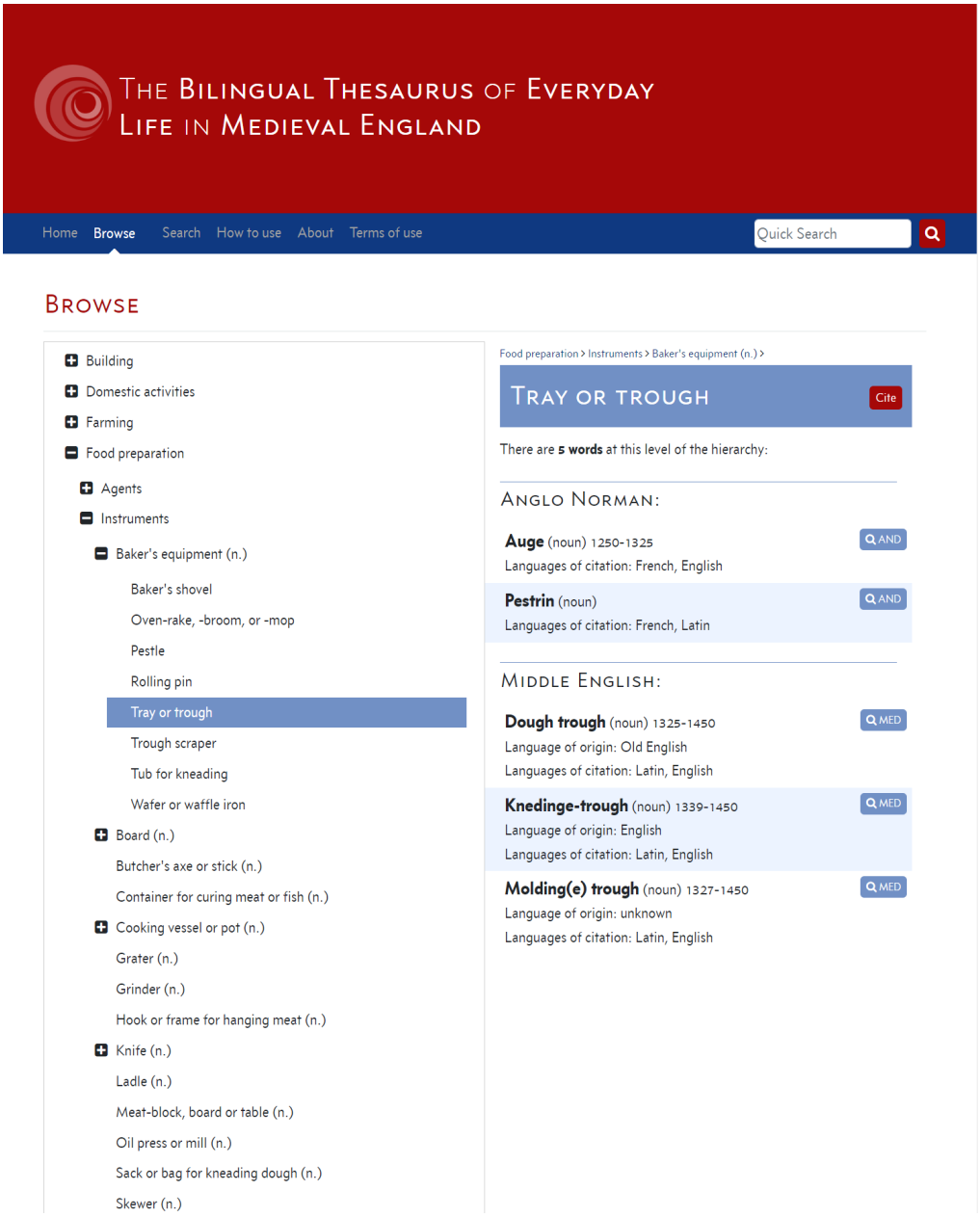


Figure 1.A.21.: *BTH* thesaurus, category “Tray or trough”.

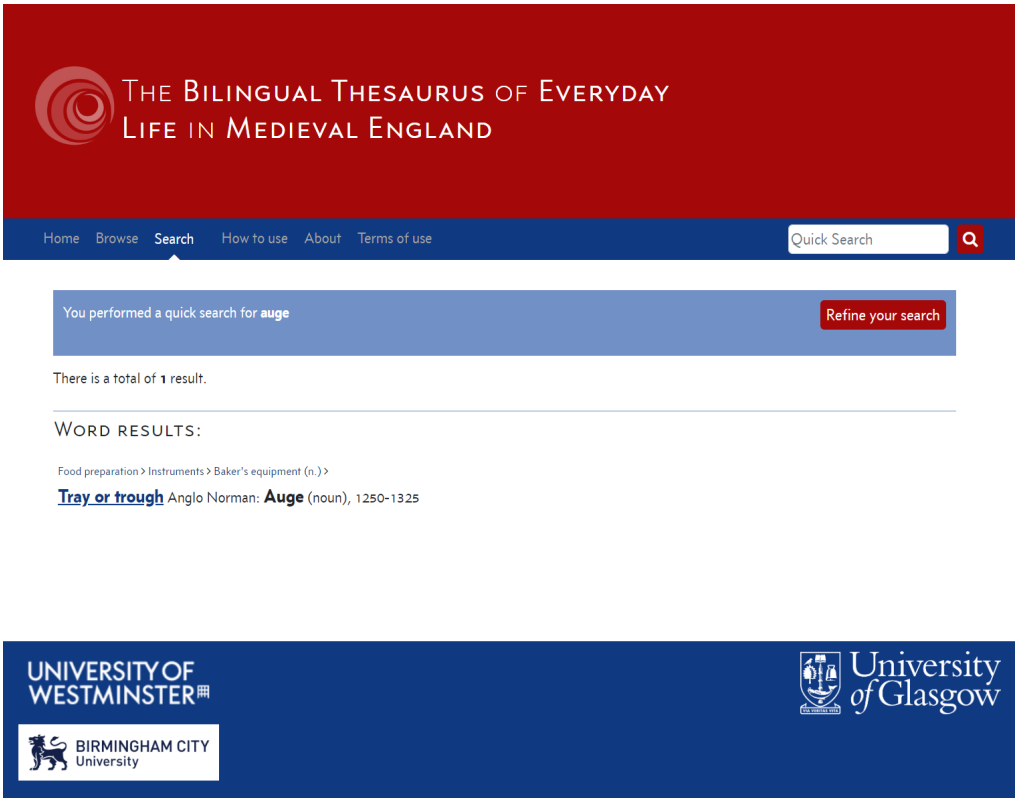


Figure 1.A.22.: *BTH* index, 'auge'.

Appendix 1.B: Usage features in historical language thesauri

This appendix details which usage features are found in the various historical language thesauri analysed in Chapter 1. First, the practice of diasystematic labelling of such features is discussed, including a typology of usage features. Afterwards, the application of each feature from this typology is considered, in separate sections, in the context of the historical language thesauri.

Diasystematic labelling

Usage features in thesauri and other lexicographical works tend to be conveyed through labelling, or rather diasystematic labelling. In these cases a label attached to a lexical sense is used to convey restrictions to which its use is subject.[85] A given sense can, for instance, be labelled *poetic* if it is restricted to use in poetry. Representing a label by an image or icon rather than by text (such as an anchor for nautical terms) is called iconic labelling.[86]

Not only the presence of a given label may convey usage information to users — the absence of a label may be equally meaningful. That is to say, labels are not uncommonly used to mark an element, implying that it "deviates in a certain respect from the main bulk of items described".[87] The absence of such a label, then, indicates that the item in question does *not* deviate from this unmarked centre in terms of its usage. Thus, omitting the *poetic* label for a specific sense may imply that it is found in prose. What is considered the unmarked centre and what is the periphery that is marked by labels can differ per lexicographical work.[88]

The decisions and implications of the labelling systems in place are, according to a number of scholars, often not described with sufficient detail.[89] In fact, some of the thesauri discussed here do not include a complete overview of all the labels in use and instead list the employed abbreviations only, which includes a subset of the labels found.[90] As Norri has shown, such lack of documentation is not uncommon for lexicographical works.[91] Indeed, the various editions of the *OED*, too, forego of providing "a list of status labels, or an explanation of how they are applied and what they mean".[92]

The labelling systems employed in dictionaries and thesauri have been subject to study, resulting in possible classifications for sets of labels that indicate a certain characteristic.[93] Of these classifications, the one fashioned by Franz Josef

[85] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diasystematic labelling'.
[86] Svensén, *A Handbook of Lexicography*, p. 319.
[87] Ibid., p. 315.
[88] Ibid., p. 316.
[89] Brewer, 'Labelling and Metalanguage', p. 493.
[90] The thesauri that do not list all labels but only the abbreviated ones are *ScT* and *LSM*.
[91] Norri, 'Regional Labels in Some British and American Dictionaries'.
[92] Brewer, 'Authority and Personality in the *Oxford English Dictionary*', p. 263.
[93] For further information on the studies and their resulting classifications, see Brewer, 'Labelling and Metalanguage', p. 498.

Hausmann appears to be the most detailed.[94] This classification provides a clear and useful framework that will be employed in this section to discuss the usage feature information found in the historical language thesauri.[95] Table 1.B.1 below presents an overview of Hausmann's typology in the context of a contemporary general-purpose dictionary.

	Criterion	**Type of marking**	**Unmarked centre**	**Marked periphery**	**Examples of labels**
1	Time	diachronic	contemporary language	archaism–neologism	*arch.*, *dated*, *old use*
2	Place	diatopic	standard language	regionalism, dialect word	*AmE*, *Scot.*, *dial.*
3	Nationality	diaintegrative	native word	foreign word	*Lat.*, *Fr.*
4	Medium	diamedial	neutral	spoken–written	*colloq.*, *spoken*
5	Socio-cultural	diastratic	neutral	sociolects	*pop.*, *slang*, *vulgar*
6	Formality	diaphasic	neutral	formal–informal	*fml*, *infml*
7	Text type	diatextual	neutral	poetic, literary, journalese	*poet.*, *lit.*
8	Technicality	diatechnical	general language	technical language	*Geogr.*, *Mil.*, *Biol.*, *Mus.*
9	Frequency	diafrequential	common	rare	*rare*, *occas.*
10	Attitude	diaevaluative	neutral	connoted	*derog.*, *iron.*, *euphem.*
11	Normativity	dianormative	correct	incorrect	*non-standard*

Table 1.B.1.: Diasystematic marking in a contemporary general-purpose dictionary.[96]

As Table 1.B.1 shows, the periphery indicated by labels used in a particular type of marking may be based on a scale (e.g., ranging from archaism to neologism, or from formal to informal). Markings based on such a scale may not be limited to just those covering diachronic, diamedial, and diaphasic usage features.[97] Another relation between labels may also exist, as pointed out by Atkins and Rundell. Diatechnical labels, for instance, may well be positioned in

[94] Vrbinc and Vrbinc, 'Diasystematic Information in the "Big Five"', p. 426. See the table on pages 428–9 of their article for a clear comparison of Hausmann's classification with other existing ones. It is worth noting that the article shows that the existing classifications, including Hausmann's, are by no means exhaustive since labels, such as 'figurative' and 'trademark', are not yet covered by them (p. 429).

[95] Hausmann, 'Die Markierung im allgemeinen einsprachigen Wörterbuch', pp. 651-2.

[96] Reproduced from Svensén, *A Handbook of Lexicography*, p. 316.

[97] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diaevaluative information', 'diafrequential information', 'diaintegrative information', 'dianormative information', 'diastratic information', 'diatextual information', and 'diatopical information'.

a hierarchical structure.[98] The labels 'Biology' and 'Astronomy', for example, could be subordinate to the more general 'Science'. Such a hierarchy has two benefits for editors: it makes it easier to ensure that no required labels are omitted and allows for more accurate (but also coarser) marking of items.[99]

Although Hausmann's typology was suggested within the subject of marking, it is useful in the broader context of lexicographical information on usage as well. Its criteria and types of marking identified can be applied to such information, regardless of whether labels are employed and regardless of whether an unmarked centre exists. Table 1.B.2 provides an overview of which kinds of usage information are conveyed systematically in the historical language thesauri analysed.

Type	***TOE***	***HTE***	***BTH***	***ShT***	***ScT***	***HTS***	***LSM***	***DSSPIEL***
Diachronic	-	+	+	-	+	-	+	-
Diatopic	-	+	-	-	+	-	+	-
Diaintegrative	-	+	-	-	-	-	+	-
Diamedial	-	+	-	-	+	-	+	-
Diastratic	-	+	-	-	+	-	+	-
Diaphasic	-	-[100]	-	-	+	-	-	-
Diatextual	+	+	-	-	+	-	+	-
Diatechnical	-	+	-	-	+	-	+	-
Diafrequential	+	+	-	-	+	-	+	-
Diaevaluative	-	+	-	-	+	-	+	-
Dianormative	+	+	-	-	-	-	+	-

Table 1.B.2.: Overview of usage information present in historical language thesauri.

As the overview in Table 1.B.2 shows, many different kinds of usage features are systematically incorporated into historical language thesauri. Most features are indicated through systematic labelling. What a label indicates can differ from one thesaurus to another. A case in point is the 'poetic' label, which in some thesauri signals that a word has a poetic flavour (e.g., *HTE*) but in others that a word is found solely in poetic works (e.g., *TOE*). Moreover, a label may encompass a greater or smaller range (e.g., for what is still considered 'formal' or 'informal') or even mark word forms or lexemes rather than senses (such as with the distribution flags found in *TOE*).[101] A label should therefore always be seen within the defined context of its body, be it a dictionary or thesaurus, and may or may not overlap with use of that same label in a different body.

Labels based on the same criterion can be related in a number of ways. Some are based on a scale; others can be positioned in a hierarchy. More intricate relations exist as well, as evidenced by diatopic usage features. The regions

[98] Atkins and Rundell, *The Oxford Guide to Practical Lexicography*, pp. 184.
[99] Ibid.
[100] Diaphasic information, at least in terms of 'formal' and 'informal' labels, is not present in *HTE3* apart from what can only be perceived as a stray marking of a single sense as 'formal': *ponor* in category "01.01.04.04.02.02|11 Land :: Hole/pit :: pot-hole/swallow-hole".
[101] Atkins and Rundell, *The Oxford Guide to Practical Lexicography*, pp. 182–6.

indicated by diatopic labels may be related to each other in various ways.[102] One region may, for example, contain, touch, or overlap with another. The regions indicated by labels are presented most vividly by a map, as is done in *ScT*, which allows the user to deduce the spatial relations between them.[103]

Diachronic information

> A usage feature which associates a word or phrase with a particular PERIOD in the history of a language. Such information can be marked in dictionaries by temporal USAGE LABELS on a chronological scale from 'archaic' through 'obsolescent' to contemporary (the unmarked neutral, synchronic zone) and 'new'.[104]

The currency of lexical senses are known in historical language thesauri. The senses found in *TOE*, for instance, belong to the Old English lexis, spoken between roughly 500 and 1100 A.D. by the Anglo-Saxons. Explicit diachronic information that further subdivides the period treated is not available in every historical language thesaurus. In fact, *TOE* treats its items as "a single geographically and temporally indistinguishable mass", ignoring diachronic differences between, e.g., Early West Saxon and Late West Saxon.[105] The currency of senses, then, is left implicit to be either within or overlap with the period specific to the thesaurus. Such a lack of explicit, finer-grained diachronic information on the currency of senses also holds true for *HTS* (or at least, for the pilot version available at the time of writing), *ShT*, and *DSSPIEL*.

Some of the historical language thesauri carry a finer-grained indication of the currency of their lexical items. *ScT* labels its senses as obsolete (with †) for cases where, as its editor states, "we have no evidence [for their use] in the twentieth century".[106] This system divides the currency of lexical senses effectively into two periods: those in use only before the twentieth century and those used also after the turn of that century. A yet greater distinction on currency than the aforementioned, relatively crude, labelling categorisation employed by *ScT* is present in *HTE*. Each lexical sense in this thesaurus "is accompanied by its dates of recorded use".[107] This information allows users to focus on the English vocabulary available in a specific time frame, such as that available to Shakespeare.[108] *LSM* and *BTH*, too provide such citation dates for the currency of their items.

Diatopic information

> A usage feature which associates a word or phrase with a particular DIALECT or regional language variety. Such features can be marked

[102] Randell et al., 'A Spatial Logic Based on Regions and Connection'.
[103] *ScT*, pp. xvii-xviii.
[104] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diachronic information'.
[105] Dance, Review of *TOE1*, p. 313.
[106] *ScT*, p. xv.
[107] *HTE1*, p. ix.
[108] *HTE1*, p. xiv.

> in dictionaries by USAGE LABELS on a continuum of regionality from 'local' or 'provincial' dialects to 'metropolitan' and even 'international' varieties. The neutral zone of the 'home' variety (e.g. British English in a British dictionary or American English in an American dictionary) may be left unmarked.[109]

Three of the historical language thesauri contain diatopic information: *ScT*, *HTE*, and *LSM*. Each of these contain the label 'dialectal' (abbreviated to 'dial' or 'dial.').[110] Next to this particular marking of dialectal items, these three thesauri may also attribute a specific region to lexical senses through labels. *HTE* and *LSM* share a number of such regional labels, including Australian (abbreviated to 'Aus' or 'Austral'), New Zealand ('NZ'), and United States ('US'). *ScT*, focused on the Scottish vocabulary, contains senses that are not as widely dispersed in use as those in *HTE* and *LSM*. Its diatopic information is, perhaps for that very reason, more specialized in nature and conveys a number of levels of granularity. The main dialect divisions which this thesaurus distinguishes for the Scottish lexis are Northern, North-East, Central, and Southern Scots, with a subdivision of Central Scots into East Central, West Central, and South-West Scots.[111] More precise regional distinctions are provided by labels representing pre-1975 counties, such as Shetland, Aberdeen, Edinburgh, and Selkirk.[112] These regional divisions are clarified visually through maps.[113] In short, the diatopic information in the thesauri is provided by labels indicating regions — regions that may themselves be subdivided into smaller identifiable regions for increased specificity.

Diaintegrative information

> A usage feature which associates a word or phrase with a particular degree of integration into the native word-stock of the language. Such information can be marked in dictionaries by USAGE LABELS on a scale of indigenisation ranging from 'foreign' and 'borrowed' through 'assimilated' to native (the unmarked neutral zone).[114]

Two of the historical language thesauri contain diaintegrative information: *LSM* and *HTE*. These thesauri mark senses of lexemes that have been borrowed from other languages and are not yet considered assimilated or native. To illustrate, both thesauri label *shadchan* as 'Jewish' and *nakodo* as 'Japanese', each in the sense of "one who makes matches". Unfortunately, the labels appear to be used rather inconsistently. Thus, *HTE* labels *kamikaze* as 'Japanese' for some senses, but not for others. In *LSM*, *maîtresse en titre* is not labelled as French but

[109] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diatopic(al) information'.
[110] Abbreviations, including those of labels, can be found on p. xxxi of *HTE1* and pp. 33-5 of *LSM*.
[111] *ScT*, p. xvii.
[112] *ScT*, p. xv.
[113] *ScT*, p. xvii.
[114] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diaintegrative information'.

instead as 'unnaturalized' (abbreviated as 'unnat') — a label seemingly just as suitable for, but not applied to, *shadchan* and *nakodo*.

In a multilingual thesaurus, language labels tend not to convey diaintegrative information but the language of its words and phrases (cf. *DSSPIEL* and *BTH*). The distinguishing factor between the two uses of these labels – i.e., diaintegrative usage versus language – is whether a lexical item is used solely in the context of a certain language or used within another language context. That is to say, if a word is employed frequently enough in a native rather than a foreign context, then the loan word may be considered part of the native lexis, albeit perhaps with its foreign origins still recognisable. Monolingual lexicographical works will, by their very definition, consider only a single language and therefore adopt these labels to portray diaintegrative usage information.

Diamedial information

> A usage feature which associates a word or phrase with a particular channel of communication. Such information can be marked in some dictionaries by USAGE LABELS for the 'written' or 'spoken' media, while items shared by both are usually left unmarked.[115]

Although none of the thesauri treated here employ the labels 'written' and 'spoken', a number of them do mark lexical senses with another diamedial label: 'colloquial'. This label is employed by *ScT*, *HTE*, and *LSM* (abbreviated to 'colloq') for terms used in daily speech.

Diastratic information

> A usage feature which associates a word or phrase with a particular social group. Such information can be marked in dictionaries by USAGE LABELS on a scale from neutral (the unmarked zone) to 'demotic' or 'slang'. In some cultures, such as the Indian caste system, there may be an extension of the scale towards the 'high' varieties. There is often an overlap with DIAPHASIC INFORMATION.[116]

Three of the historical language thesauri contain diastratic information: *LSM*, *HTE*, and *ScT*.[117] All of these contain the label 'slang' (abbreviated in *LSM* to 'sl'). *HTE* describes the meaning of this label as one denoting "very informal language that is often restricted to a particular social group".[118] Labels indicating a specific form of slang, namely rhyming slang, are found in *LSM* (where it is abbreviated to 'Rsl'). One instance for which this label is applied is the phrase *trouble and strife* in the sense of wife.[119] In addition to 'slang', both *LSM* and

[115] Ibid., s.v. 'diamedial information'.

[116] Ibid., s.v. 'diastratic information'.

[117] Although the use of the label 'slang' is not indicated in the introduction of *ScT*, it is certainly employed systematically as evidenced by the lexical sense of *bung* in category "2.4 Horses, donkeys".

[118] *HTE1*, p. xxiv.

[119] See *LSM* category "M.01.01.01 Wife".

HTE employ the diastratic label 'vulgar' (abbreviated to 'vulg' in *LSM*) for lexical senses that are considered rude or obscene.[120]

Diaphasic information

> A usage feature which associates a word or phrase with a particular REGISTER of a language. Such information can be marked in dictionaries by USAGE LABELS on a scale from 'elevated' and 'formal' through neutral (the unmarked zone) to 'informal' and 'intimate'. There is often an overlap with DIASTRATIC INFORMATION.[121]

Diaphasic information is conveyed in *ScT* through the label 'informal'.[122] *HTE* makes no mention of that specific label in its front matter, but contains what appears to be a single stray application of the label 'formal'.[123] Diaphasic markings do not seem to be present in the other thesauri analysed in the chapter.

Diatextual information

> A usage feature which associates a word or phrase with a particular discourse type or GENRE. Such information can be marked in dictionaries by USAGE LABELS on a scale of textuality from 'poetic' to 'conversational', with the shared neutral items remaining unmarked. There is often an overlap or combination with DIAPHASIC and DIASTRATIC information.[124]

The historical language thesauri analysed that employ usage labels all include diatextual information. One example is the restricted use to poetry or verse, which is indicated by a label in *TOE* (with the *p*-flag), *HTE* and *LSM* (with the abbreviated label 'poet') as well as *ScT* (using the label 'verse'). A second example is the label for glossaries that is applied in *TOE* (with the *g*-flag) and *LSM* (with the abbreviation 'gloss'). However, as the editor of *LSM* notes, identical diatextual labels found in different thesauri may not have the same meaning that they have at first glance: contrary to the practice in *LSM*, the editors of *TOE* employ the glossary label to "flag terms occurring only in glossaries, but not terms found in glosses".[125] Moreover, the labels in *TOE* "relate only to word forms, not to meaning".[126] In other words, even though they may be presented at every lexical sense, these labels do not indicate usage features

[120] *LSM*, p. 35; *HTE1*, p. xxiv.
[121] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diaphasic information'.
[122] The marking 'informal' in *ScT* is, for instance, attributed to *Royal* in category "12.2 Soldiers and other services personnel".
[123] The marking 'formal' in *HTE3* can be found for the lexical sense *ponor* in category "01.01.04.04.02.02|11. pot-hole/swallow-hole".
[124] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diatextual information'.
[125] *LSM*, p. 34.
[126] *TOE2*, p. xxi.

of each sense separately. The editors of *TOE* state that such labelling per sense will demand further efforts in Old English lexicography.[127]

Diatechnical information

> A usage feature which associates a word or phrase with a particular SUBJECT FIELD. Such information can be marked in dictionaries by USAGE LABELS for a range of technical specialities, e.g. 'Law', 'Music', 'Chemistry'. The CORE WORDS common to non-technical language varieties are usually left unmarked, as is non-specific, 'vague' vocabulary.[128]

ScT contains at least one label referring to a subject field: archaeology.[129] *HTE* and *LSM* both contain several diatechnical labels, a number of which are shared; both draw from the *OED*, which includes such markings, as their source. Subject fields indicated by labels in the two thesauri include archaeology, nautical, military, theatre, and botany.

Diafrequential information

> A usage feature which associates a word or phrase with a particular FREQUENCY of occurrence. Such information can be marked in dictionaries by USAGE LABELS ranging from 'very frequent' to frequent (the unmarked neutral zone) to 'becoming rare' and 'very rare'.[130]

Four of the historical language thesauri treated here annotate their lexical senses with diafrequential information. Three of them – *ScT*, *LSM*, and *HTE* – mark senses when they are rare. Extremely rare items, or nonce words, are marked in *TOE*, *LSM*, and *HTE*. Such items occur only once in the corpus that the thesaurus (directly or indirectly) is based on. *TOE* employs the *o*-flag for this marking; *LSM* and *HTE* use the label 'nonce' and 'nonce word' respectively.[131] The editors of these works are aware that designating a word form in a specific sense as a nonce word, or *hapax legomenon*, is fraught with problems.[132] Do derivations of the word found in another part of speech negate the status of nonce? Does evidence in a later phase of the language than that treated in the thesaurus entail that a word should not be marked as nonce? If the corpus is expanded, is it not likely that the nonce status may disappear? As the editors of *TOE* state, such labelling of nonce words "as often as not should serve to give rise to speculation and inquiry".[133]

---

[127] textitTOE2, p. xxii.
[128] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diatechnical information'.
[129] This label is never found as 'archeol', the abbreviation listed in the introduction, but rather as 'archaeol'. See, for instance, *yett* in category "13.3.5 Castles etc.".
[130] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diafrequential information'.
[131] *TOE2*, pp. xxi–xxv; *LSM*, p. 34; *HTE1*, p. xxiv.
[132] *TOE2*, p. xxi.
[133] *TOE2*, p. xxv.

Diaevaluative information

> A usage feature which associates a word or phrase with a particular attitude or evaluation. Such information can be marked in dictionaries by USAGE LABELS on a scale of emotiveness from 'appreciative' through neutral (the unmarked zone) to 'derogatory' and 'offensive'.[134]

Diaevulative information can be found in the form of labels in *ScT*, *HTE*, and *LSM*. All three employ labels to mark derogatory and jocular lexical senses (abbreviated to 'derog' and 'joc', respectively). *HTE* also employs a label to indicate contemptuous use for some of its senses (abbreviated to 'contemp'). Although some of the contemptuous lexical senses found in *HTE* are also present in *LSM*, it appears that what is marked as contemptuous in the former is simply labelled derogatory in the latter.[135]

Dianormative information

> A usage feature which associates a word or phrase with a particular degree of deviation from a cultural STANDARD. Such information can be marked in dictionaries by USAGE LABELS on a scale from correct (the unmarked neutral zone) to 'substandard' or 'illiterate'.[136]

Dianormative information is found in a number of the historical language thesauri. *TOE*, for instance, employs the *q*-flag for lexical senses that were deemed questionable as to whether they have truly existed in Old English.[137] Questionable forms may be those postulated and altogether absent from the extant corpus of Old English or have been deemed to likely be misreadings. Their retention rather than exclusion in the thesaurus is "a matter almost of book-keeping", underlining that there is a lack of evidence for these lexical senses found in the source dictionaries of the Old English language.[138]

ScT is somewhat enigmatic in its approach towards dianormative information. The list of abbreviations, part of its front matter, includes 'St' for standard and 'erron' for erroneously. With the lexical senses themselves, however, the former abbreviation is used only for Saints (e.g., St Andrew) and not to represent whether a lexical sense is deemed standardized. The abbreviation 'erron' for erroneously, too, is nowhere to be found in the content proper. Although the source dictionary for *ScT* applied this label for *politik*, for instance, *ScT* has dropped its marking altogether.[139] *ScT*, then, does not systematically contain dianormative information. Its listing of the aforementioned two abbreviations,

---

[134] Hartmann and James, *Dictionary of Lexicography*, s.v. 'diaevaluative information'.

[135] See *puppy-love*, which is found in category "02.04.13.11|16 romantic attachment between boy and girl" in *HTE3* and in category "L.04/01.01.01.01 Puppy love" in *LSM*.

[136] Hartmann and James, *Dictionary of Lexicography*, s.v. 'dianormative information'.

[137] *TOE2*, pp. xxx-xxxi.

[138] *TOE2*, p. xxx.

[139] *ScT: politik* in category "15.2.1 High class". Cf. the corresponding sense in the source dictionary: *The Concise Scots Dictionary*, s.v. 'politik', sense 2.

which are never employed in the thesaurus, confirms the claim made by Brewer and others that usage labelling is far from always dealt with in a consistent manner.[140]

Both *HTE* and *LSM* employ the label 'erroneous' (abbreviated to 'erron') to mark incorrect use of a word — either because the word itself is deemed a malformation or the sense attributed to it is considered flawed. Remarkably, even though both thesauri have incorporated data from the *OED*, they may differ in which senses they actually mark as erroneous. For example, *moryeve < morgengifu* in the sense of 'dowry' is marked in *HTE* as erroneous; in *LSM* it is not.[141] The same holds for *patriot (of)* in the sense of 'one who loves'.[142] Moreover, some of the senses marked as erroneous in *LSM* are not found, or at least not in those attributed senses, in *HTE*. Examples of such items are *agapemone* in the sense of 'a love feast', and *oscultation* (rather than *osculation*, which is listed at the same location) in the sense of 'kissing'.[143] Additionally, *LSM* employs a dianormative label not found in *HTE*: 'dubious'. The term itself is not explained in the thesaurus, but the name suggests it might be considered to convey a judgement similar to 'erroneous', albeit perhaps a weaker variant. An example of this label in use is in the marking of *lovely* in the sense of 'inspired by love' — a sense that is never attributed to *lovely* in *HTE* and will not have been found in the *OED* but elsewhere.[144]

The electronic edition of *HTE* incorporated into *OED Online*, *HTE2*, will certainly see updates in its dianormative as well as other types of labelling owing to the progress towards a third edition of this dictionary. The editors of the *OED* appear to move from prescriptive judgements on lexical senses to "tackling issues of usage and correctness with informed impartiality" by quoting from prescriptive sources rather than taking a stance on the matter itself.[145] Although the third edition of the *OED* has yet to be completed, Brewer's analysis of the work done so far suggests that the "most significant changes must, surely, be the avoidance of the terms 'erroneous' and 'catachrestic' [...] and the decision not to use the paragraph mark as an indicator of incorrect usage".[146]

[140] Brewer, 'Labelling and metalanguage', pp. 488–9; Norri, 'Regional Labels in Some British and American Dictionaries', p. 26.

[141] *HTE3*, s.v. 'moryeve < morgengifu' in category "03.01.01.05.11.08|03 Gifts and payments :: dowry". *LSM*, s.v. 'moryeve < morgengifu' in category "M.04/11.03.01.01 Dowry :: woman's marriage portion (in money/goods)".

[142] *HTE3*, s.v. 'patriot (of)' in category "02.04.13|03 Love :: one who loves". *LSM*, s.v. 'patriot (of)' in category "L Love :: ..one who loves".

[143] *LSM*, s.v. 'agapemone' in category "L/01.03 Spiritual Love :: ..a love feast (like those) held by the early Christians". *LSM*, s.v. 'oscultation' in category "L.01.02.02.02 Kissing :: the act of touching with the lips in token of endearment". These two senses were not present in the *OED*, which suggests Coleman must have taken these from other sources. It would also explain why these do not occur in *HTE*, since that thesaurus has used only the *OED* as its source for the English vocabulary more recent than Old English.

[144] *LSM*, s.v. 'lovely' in category "L/02 Aspects of Love :: ...inspired by love". This lexical sense was not present in the *OED*. The implication is alike to that discussed in footnote 143.

[145] Brewer, 'Prescriptivism and Descriptivism in the First, Second and Third Editions of *OED*', p. 31.

[146] Brewer, 'Authority and Personality in the *Oxford English Dictionary*', p. 299.

Chapter 2

2. Functionalities required of historical language thesauri for research and education

In order to improve the dissemination of historical language thesauri on the Web, not only is their content of importance but also the functionality required by their target audience. This chapter describes functionality required by users in academia, both for research and educational purposes, of editions of these lexicographic resources. In order to gather information on such functionality, a number of sources have been consulted. First, existing editions of historical language thesauri and handbooks on both thesauri and lexicography in general, which were also consulted in Chapter 1, provided valuable input. Second, academic reviews of these thesauri and notable research employing these resources offer further insights. Lastly, additional input has been collected from stakeholders – experts in lexicography, linguistics and philology (amongst other fields) – through dedicated stakeholder meetings, workshops, and feedback based on preliminary results in research and education in a research project titled 'Exploring Early Medieval English Eloquence'.[1] Within the remainder of this chapter, the present set of functionality gathered is detailed in five sections: navigation (2.1), resource views (2.2), extension (2.3), analyses (2.4), and data management (2.5).

2.1. Navigation

The first and foremost functionality required of thesaurus editions is the ability for users to navigate their content. The preface to *HTE* indicates that users have a choice in how to approach that content:

> There are two ways to approach a thesaurus: by familiarizing oneself with its structure and principles of organization, or, more commonly, by using its index to determine in which category or categories a word appears.[2]

Whether the one approach is truly used more commonly than another is less clear cut. Christian Kay, one of the editors of *HTE*, mentions that the alphabetical index is there "for convenience" and may not be needed for "confident and

[1] The research project and its resulting case studies are discussed in Chapter 8. Appendix 2.A lists the functionality required by researchers participating in this project.

[2] *HTE1*, p. ix.

frequent users" of the thesaurus.[3] Similarly, the editors of *TOE* prompt the user to approach their thesaurus "by subject rather than through the alphabetic index".[4] Nevertheless, both means of navigation are offered to users of historical language thesauri and hence warrant further treatment.

2.1.1. Navigation via the thesaurus structure

Users can employ the topical structure of a thesaurus to move from meaning to lexical items conveying that meaning. In order to facilitate this approach, print editions of the historical language thesauri under discussion all include an outline of their topical system.[5] For those historical language thesauri that have three or fewer levels in their topical system, such as *ScT*, this outline typically presents the system in its entirety. For those thesauri with a larger degree of specialization in their topical systems, termed distinctive thesauri by Christian Kay and Marc Alexander, the outline may be limited to the top categories in order to save space.[6] Such a limited outline is offered in the print editions of *TOE* and *HTE*, which contain over ten levels in their topical systems and provide an outline for the first two and three of these levels, respectively.

Through the use of hyperlinks, electronic editions of historical language thesauri can supply users with the means to navigate the topical system in an interactive manner. The majority of the current electronic editions of these thesauri offer an overview similar to those found in the print editions. The overviews of *BTH*, *TOE4*, *HTE2*, and *HTE3* are located on a single webpage, allowing the user to expand or collapse categories to show or hide subordinate ones. Only when selecting a specific category for viewing purposes will this edition show the senses contained within. Such a useful overview has not been offered by all Web-based editions of thesauri. *TOE3*, no longer available, and *HTS* require the user to navigate the structure by taking single steps further down, or up, the hierarchy. A new webpage is loaded in the browser for each position in the topical system that the user passes through. In contrast, all electronic editions indicate the current position in the taxonomy through breadcrumbs, a trail of hyperlinks of the current category and ones superordinate to it.

2.1.2. Navigation via the thesaurus index

The second manner in which users can approach a thesaurus is through an alphabetic index. Reviewers of historical language thesauri have called the availability of such an index "necessary" and even "indispensable" in order to find a word or phrase within the categories of thesauri.[7] The added value of

[3] Kay and Alexander, 'Diachronic and Synchronic Thesauruses', p. 368.
[4] *TOE2*, p. xv.
[5] *TOE2*, pp. v–viii; *LSM*, pp. iv–xvii; *HTE1*, pp. xxix–xxx; *ShT*, pp. xvii–xxv; *ScT*, pp. iv–vii. *DSSPIEL* has separated its outline over the table of contents for the primary categories (p. xix) and the sections that treat each of those categories. These sections begin with a list of the subcategories.
[6] Kay and Alexander, 'Diachronic and Synchronic Thesauruses', p. 370.
[7] Görlach, Review of *TOE1*, p. 399; Momma, Review of *TOE2*, p. 80.

an index seems to have been well understood by the editors of the historical language thesauri of Scots and English, since all their printed editions include an alphabetic index to their contents.

Which parts of the contents are indexed alphabetically can vary between thesauri in print. The majority of the historical language thesauri treated here – *ShT*, *LSM*, *TOE*, and *HTE* – index their lexical senses (through their headforms), but not their categories. The editors of the other two printed historical language thesauri have opted to base their index on English rather than their actual lexicon. As such, the index that complements *DSSPIEL* is one to its categories, or headings, rather than to its categorized lexical items.[8] The index of *ScT* is meant to provide "a detailed signposting in English into the intricacies of Scots vocabulary",[9] listing near English equivalents to the Scottish senses, when available, as well as more generic English concepts taken from its English categories.[10] In other words, this index incorporates elements from both the lexical senses (i.e., their definition) and the topical system in order to assist the reader, who is likely to learn more about Scots through the English language than through the Scots vocabulary directly.

Although they have their use for thesauri in printed from, alphabetic indexes are no longer essential for digital thesauri. Instead, the functionality required to find a word or phrase may be provided by a search system. Indeed, the digital editions of *TOE* and *HTE* sport such search systems and have abandoned an alphabetical index. These search systems offer a vast improvement over an index in three ways. The first is in terms of retrieval speed — going back and forth between index and the thesaurus proper can now indeed take only "a twinkling of an eye".[11] The second improvement is that both categories and lexical senses can be included in searches without any serious drawbacks in this method of access. Where textual editions were criticized for not containing alphabetical indexes to both categories and entries,[12] possibly due to limits on printing space imposed by publishers or to avoid too great a weight of the books for readers to handle them, the digital editions of *TOE* and *HTE* offer this sought-after ease of access through their new search systems. The third improvement offered by digital search systems is that lexical senses can now be located based on the additional information per sense, including the labels they carry. This allows searches to be restricted, for instance, to items of a particular part of speech or to those marked as 'poetical'. Of course, it would be just as interesting to restrict searches to senses that do *not* carry a given label, since these senses are implicitly marked by the editors as belonging outside of the group indicated by the absent label.[13] The possibility to exclude a label from searches would allow one, for example, to locate only non-poetical lexical senses. This functionality of exclusion of labels in searches is not yet provided by the current systems in

[8] *DSSPIEL*, p. 1505.
[9] *ScT*, p. xix.
[10] *ScT*, p. xix.
[11] Busse, 'A Celebration of Words and Ideas', p. 804.
[12] Görlach, Review of *TOE1*, p. 399.
[13] Svensén, *A Handbook of Lexicography*, p. 315.

place for the digital versions of *TOE* and *HTE*.

2.2. Resource views

The second functionality required of thesaurus editions is the ability to view all the information available on a specific category or lexical item that a user chooses to inspect. For lexical items these resource views ideally contain the part of speech, language, and usage features; and for categories, their name and identification string.[14] Moreover, these overviews should indicate relations to other resources where relevant, such as, for categories, a list of the lexical senses that are allocated to the viewed location of the thesaurus taxonomy or, concerning lexical items, an overview of synonyms. These essential resource views are present in historical language thesauri (e.g., as entries in a printed thesaurus), albeit not always as complete as possible.

One piece of information useful for research, noted by both editors and reviewers of thesauri and absent from existing resource views of lexical items, is an overview of the different senses of a single lexeme.[15] Such views of senses related through polysemy facilitate explorations of metaphorical and metonimical uses and offer a measure of ambiguity in utterances in which the lexeme in question occurs. Through these connections of individual lexemes, it is possible to detect which semantic fields in a language have close conceptual ties. For example, terminology surrounding sleep can be used metaphorically for death; temperature for emotions.[16] By mapping out metaphorical connections such as these, it is possible to gain a better understanding of the stylistic impact of metaphors and to grasp which groups of words are more easily used to symbolically represent other meanings.[17] Existing historical language thesauri of Scots and English, rather than including the information on senses related through polysemy in the resource view of a lexical item, rely on the user switching to either the index (in print editions) or the search engine (in digital editions) to obtain this knowledge on the lexical item inspected.

Resource views could also be employed to present details on elements other than categories and lexical items. Labels found in thesauri are valuable elements of information, too, that may warrant their own resource view. Such labels, which benefit from thorough descriptions, are utilized by researchers to investigate the usage features with which those labels are associated.[18] Providing resource views for these elements could assist researchers in finding helpful information on them, including a full name and description instead of only a code or abbreviation. Moreover, providing these resource views for labels can assist in retrieving the elements they mark. One such need is exemplified in Kathryn Allan's work on

[14] See Chapter 1 for a thorough overview of information available in historical language thesauri.
[15] See *HTE1*, p. x; Ilson, 'On the Historical Thesaurus of the Oxford English Dictionary', p. 256.
[16] For both metaphorical ties mentioned, see *Mapping Metaphor with the Historical Thesaurus*.
[17] Ibid.
[18] Brewer, 'Labelling and Metalanguage', p. 493; Norri, 'Regional Labels in Some British and American Dictionaries'. See the discussion in section 1.8 of Chapter 1.

metaphor and metonymy, drawing on information from *HTE*.[19] Lexical senses in *HTE* were marked with labels that represent what Allan calls core concepts. Subsequently these custom labels were used to retrieve all items tagged with a specific core concept, such as ANIMAL or AGE.[20] Although the need for adding custom labels will be explored further in section 2.3, the demand for viewing a list of all items to which a label is applied can be satisfied through a resource view for this kind of element. Despite these reasons, none of the Web-based editions of historical language thesauri offer resource views for elements other than categories and lexical items.

Lastly, resource views would benefit from the means to hide information deemed irrelevant to a user. This aspect has been noted by creators and reviewers of thesauri alike as important.[21] The editors of *HTE*, for instance, state that the very purpose of their thesaurus is "to provide a detailed record of the English vocabulary from the earliest times to the present, with sufficient accompanying information that, for any given period in the past, the user should be able to ascertain the exact state of the vocabulary (i.e., the 'lexical system') which existed at that time".[22] They point out that the *HTE* could thus be used to "act as a thesaurus for any period in the past", allowing one to determine which lexical items will have been available to Shakespeare.[23] However, both digital editions of *HTE* prevent the creation of a subthesaurus based on the recorded diachronic usage features, containing only those lexical senses that were available in Shakespeare's time — or any other subthesaurus based on a set of criteria for that matter. Any sifting is left as manual labour to the user. *TOE*, too, includes valuable tagging information, stating whether its items are found only in poetry or only in glosses, for example. Although *TOE3* allowed viewing subthesauri based on the available tagging information, *TOE4* no longer sports this helpful feature.[24] In short, the existing digital versions of these thesauri lack the ability for their users to view only those items deemed of interest based on usage features and other information available — a filtering ability applicable to both resource views and the mechanisms available to navigate a thesaurus.

2.3. Extension

The third functionality researchers require of thesaurus editions is the ability to extend them, connecting additional information to their content in order to enrich or reuse that existing data. Examples of such extensions are indications of date and dialect, results from corpus searches, and indications whether a word or meaning is found in a particular text, context, or is notable in some other

[19] Allan, *Metaphor and Metonymy.*

[20] Ibid., pp. 21–2.

[21] For an academic review discussing the possibility and desirability of creating subthesauri through such mechanisms of filtering, see Conner, Review of *TOE1*, p. 888.

[22] *HTE1*, p. xiii.

[23] *HTE1*, p. xiv.

[24] For a description of the previously available *TOE* website and its feature to create subthesauri based on tagging information, see Stolk, 'Welcoming the *Thesaurus of Old English Statistics*', pp. 11–14.

qualitative or quantitative way. The functionality to extend thesaurus content offers users the means to have the thesaurus reflect their own interests and to share salient information with others. This section discusses two forms of extension: 1) elaboration on thesaurus content and 2) connecting other bodies of knowledge to a thesaurus.

2.3.1. Elaboration on content

Many scholars desire labelling information in historical language thesauri beyond what has been made available and, for that purpose, the ability to elaborate on thesaurus content. In a review of *ShT*, for instance, Christian Kay expresses that she finds it "frustrating that nothing other than the part of speech of each word is offered to the user".[25] A second example is the existing tagging information per lexical item in *TOE*, which is deemed helpful and efficient,[26] but also thought to be rather limited.[27] Indications of date and dialect, for example, are notably absent. As it stands, all items are treated as belonging to "a single geographically and temporally indistinguishable mass".[28] Researchers certainly possess further insights on these topics: Rolf H. Bremmer Jr asserts that the dialectal origins of many Old English words can be pinpointed globally as Anglian, West-Saxon, or Kentish and that such information would be worthwhile to add to *TOE*.[29] The added knowledge would facilitate research into the impact of regional influences on the Old English vocabulary as it has come down to us and, further along the line, on present-day English. Researchers participating in the workshop series, too, indicated the usefulness of adopting custom labels to add further distinctions in genre, dialect, diachronic usage, and etymology.[30]

The means to elaborate is also desired on elements other than labels for lexical senses. Kathryn Allan, for example, points out that sections of *HTE* available to her lacked lexical items that she thought belonged in these locations and, consequentially, wished to insert there.[31] Similarly, certain lexical items may warrant marking to be stricken from a thesaurus — especially if dealing with words that appear to lack evidence for their existence, so-called ghost words.[32] Another case for elaborating has been made by Nils Århammar, who suggests supplementing *DSSPIEL* with lexical items of various stages of the Frisian language.[33] Additions and refinements to the topical system, too, are valued for research purposes. Christian Kay, in her review on *ShT*, professes that she would welcome a further subdivision in its existing system of rather abstract categories. Such a fine-grained division "would make it easier to identify areas of

[25] Kay, Review of *ShT*, p. 72.
[26] Cavill, 'Names and Things in Anglo-Saxon and Early Norman England', p. 186; Conner, Review of *TOE1*, p. 888.
[27] Bremmer, 'Treasure Digging in the Old English Lexicon', p. 111; Görlach, Review of *TOE1*, p. 399; Dance, Review of *TOE1*, p. 313.
[28] Dance, Review of *TOE1*, p. 313.
[29] Bremmer, 'Treasure Digging in the Old English Lexicon', pp. 111–12.
[30] See Chapter 8 and Appendix 2.A.
[31] Allan, *Metaphor and Metonymy*, p. 20.
[32] Hartmann and James, *Dictionary of Lexicography*, s.v. 'ghost word'.
[33] Århammar, 'A Frisian Supplement to Buck's Dictionary of Indo-European Synonyms?'.

high lexicalisation [i.e., where elaboration of vocabulary is notable], and, with the help of other reference works, to examine areas where Shakespearean innovations are prevalent".[34]

A perhaps more drastic fashion in which to extend thesaurus content is by creating a new topical system to supplant the existing one, reusing and extending only lexical items recorded in a thesaurus. A case in point is Thijs Porck's study of the conceptualisation of old age in early medieval England, which draws on the contents of *TOE*, but replaces the existing categorisation with one more suitable to capture nuances newly established in the semantic field under investigation. Lexical senses from four separate *TOE* categories are grouped together as all belonging to the category of "Human old age".[35] The new categorisation system fashioned for these items contains semantic domains, such as "Positive development", "Wisdom", and "Authority". Similarly, *BTH* indicates that it reuses categories from the topical system of *HTE*, "though these were occasionally modified in order to capture the ways in which conceptions in the Middle Ages differed from those of the present day", and adds words from Anglo French, spoken in England alongside Middle English after the Norman Conquest in 1066.[36] In short, many scholars wish to extend the content of historical language thesauri, to varying degrees, for research purposes.

Unfortunately, any elaboration on thesauri content and subsequent sharing of these additions is currently not facilitated by most of the published forms of the historical language thesauri treated here. Only *HTS*, one of the thesauri available online, sports a comment section in which users can share their thoughts publicly (see Figure 2.1). All other publications of the thesauri analysed do not provide the means for additions to be made by anyone other than their editors. One can scribble additions in the margins of paper copies or refer to bits of information via a textual reference, but these acts do not extend the conceptual thesaurus: The additions either affect only a single copy (in the case of annotations in the margins) or are stored in a different location and format that hampers simultaneous and integrated access to the content of both the original and its additions.

2.3.2. Connecting bodies of knowledge

Expansion can come not only in the form of elaboration on existing thesaurus content, but also by connecting other bodies of knowledge to a thesaurus. The desire for such functionality is expressed both by scholars who utilise or review the thesaurus and the editors themselves. The introduction to the paper editions of *TOE*, for instance, states the following: "Should a precise meaning be wanted, a dictionary is needed".[37] *ScT* contains a similar statement, referring readers to the *Concise Scots Dictionary* for etymology and pronunciation of items from its lexis.[38] The introduction to *HTE*, too, indicates that connectivity between

[34] Kay, Review of *ShT*, p. 73.
[35] Porck, 'Growing Old among the Anglo-Saxons', pp. 68–9.
[36] See *BTH*, section 'About'.
[37] *TOE2*, p. xv.
[38] *ScT*, p. xv.

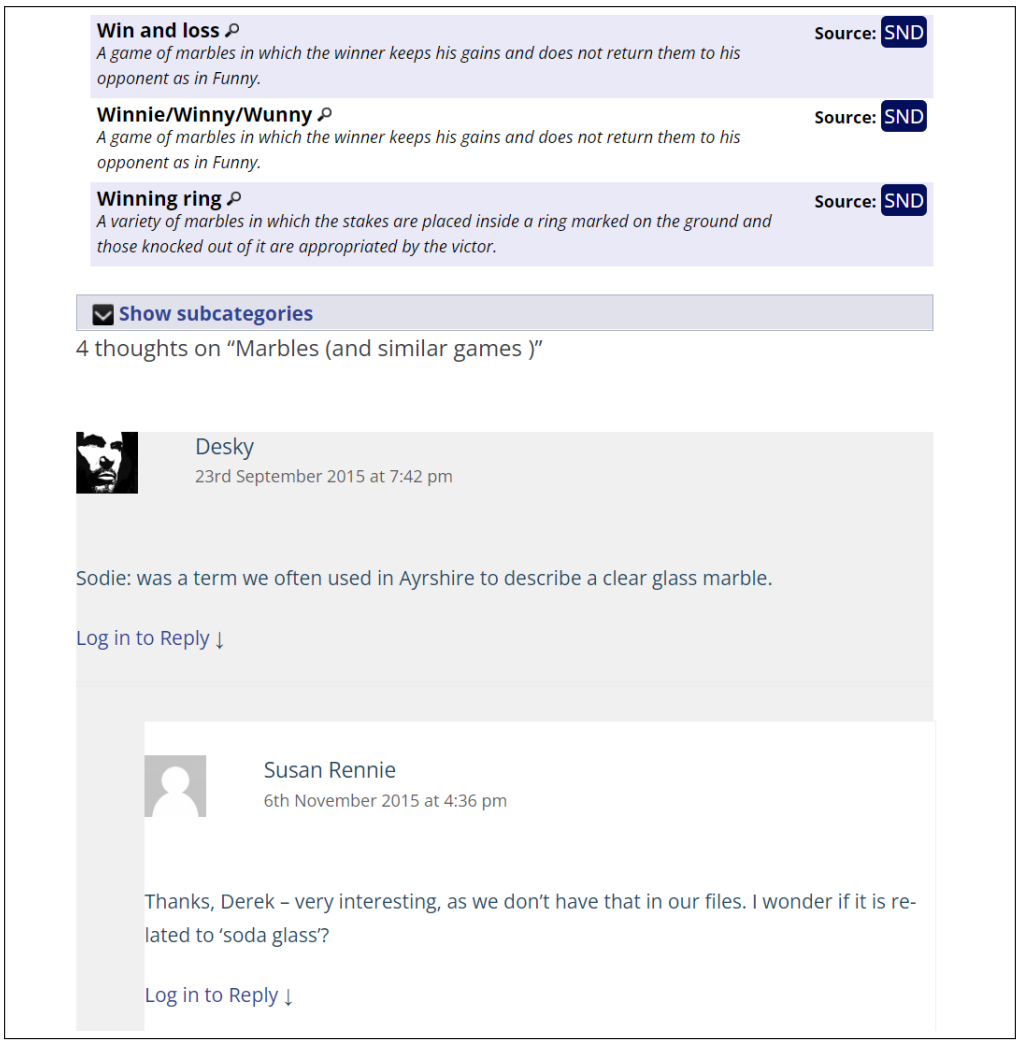


Figure 2.1.: Comments by users on *HTS* category "Marbles (and similar games)".

the thesaurus and a dictionary with supplementing information, the *Oxford English Dictionary*, is desirable.[39] That such relations between bodies are worth exploring is not only suggested by editors but also taken to heart by users. David Crystal, for one, in his explorations of *HTE* content has also turned to related definitions and citations from the *OED*.[40] Likewise, Kathryn Allan has performed analyses of *HTE* items "by looking closely at etymological information supplied in the *OED*".[41] Relations between bodies of knowledge like the aforementioned ones can, as the editor of *ScT* points out, act as "a series of step-by-step doorways into the heart of a national culture".[42]

Relating bodies of knowledge to each other can be achieved in a number of

[39] *HTE1*, p. xiv.
[40] Crystal, *Words in Time and Place*, p. xv.
[41] Allan, *Metaphor and Metonymy*, p. 21.
[42] *ScT*, p. ix.

ways. The extent to which the user is facilitated in retrieving and combining information from multiple bodies varies per solution. The most basic level of relating two bodies is by mentioning to users that such a relation exists, possibly in the introduction. This method is applied in many print editions of historical language thesauri.[43] It is wholly left to the user to access the second body of knowledge and subsequently to locate the desired complementary information. A more user-friendly solution of relating two bodies is by making an explicit mention of where exactly such complementary information can be found. External references per lexical sense do just that.

Although an external reference at a lexical sense directs users to the appropriate location in another body, paper editions still require users to move from body to body in order to accumulate the information sought after. Facilitating users in the fullest manner possible comes from connectivity in a digital environment, such as through hyperlinks. In fact, true connectivity would allow users to access and query knowledge from multiple bodies at the same time, without having to manually move from the one to the other. Unfortunately, as McCracken notes, connecting bodies is "curiously underexplored in lexicography", even though it is "a familiar topic in the context of knowledge bases".[44] By applying conventions found in the context of knowledge bases, it should therefore be possible to remove the somewhat isolated status of historical language thesauri. For users to have complementary information at their fingertips, editors and publishers of thesauri may well want to shift their focus from "aspirations of completeness and comprehensiveness" to "connectivity and interoperability".[45]

2.4. Analyses

The means to perform statistical analyses is another key feature sought after by researchers.[46] Such analyses, utilizing the onomasiological structure of the thesaurus and features of the lexis it contains, facilitate investigations into a range of aspects encoded in the lexicon: cultural elaboration, semantic domains and their cultural connotations, stylistic preferences of authors, use and development of metaphors, and so on.[47] Regrettably, few editions of historical language thesauri offer functionality for even the most rudimentary of analyses.

ShT, Christian Kay observes, sorely lacks a simple count of lexical items under a category.[48] Such counts indicate the degree of lexicalization, also known as cultural elaboration, of semantic concepts.[49] The underlying hypothesis for the

[43] See Chapter 1, section 1.4.1.
[44] McCracken, 'The Exploitation of Dictionary Data and Metadata', p. 513.
[45] Ibid., p. 513.
[46] Evidenced by Appendix 2.A, which indicates such functionality has been required by researchers in the project 'Exploring Early Medieval English Eloquence'.
[47] See, for instance, *ShT*; Crystal, *Words in Time and Place*; *Mapping English Metaphor Through Time*; Porck, 'Growing Old among the Anglo-Saxons', pp. 59–71; Diller, 'Measuring the Growth of Semantic Fields'.
[48] Kay, Review of *ShT*, p. 73.
[49] Wierzbicka, *Understanding Cultures through Their Key Words*, pp. 10-11.

importance of these figures is that domains that are important in a culture are heavily encoded in the language of that community, providing its speakers with a multitude of nuances to discuss the subject. For thesaurus content in a digital environment, this statistic is relatively straightforward to obtain. Even so, many thesauri – both paper and electronic editions – do not include these statistics in their editions, including the paper editions of *TOE* and *HTE*. However, their electronic editions *TOE3*, *TOE4*, and *HTE3* currently indicate the number of lexical senses located at any given category but, regrettably, without presenting the accumulated figure for that category and its subordinate categories, i.e., for the semantic domain.[50] The larger the semantic domains analysed, the more valuable it will be for researchers to have these statistics automatically generated as opposed to calculating them manually. In *TOE*, for instance, 1,348 lexical senses evoke the concept "13.02 War", whereas the nine co-ordinate domains for peace encompass a mere 119 lexical senses.[51]

Statistics more nuanced than a count of all items located at a specific category are absent from all editions of historical language thesauri. This absence hampers researchers in utilizing the semantic information encoded by the topical system for an onomasiological analysis of lexical items with specific characteristics. Salient information on lexis – such as period, region, register, or use by a specific author – can form the basis of investigations into their spread across the semantic hierarchy, be it horizontally over the various semantic fields or vertically between levels of specificity in meaning attributed to the lexical items selected. Thus, the hierarchy of a historical language thesaurus has the potential to act as "summary of the semantic framework" and yield onomasiological profiles for those words and phrases in which researchers are interested.[52] Additionally, researchers participating in the workshop series indicated the desire to contrast analyses between different languages and other features attributed to lexical items.[53]

2.5. Data management

The fifth functionality required for research is the ability to manage the data under investigation, a demand that is largely the consequence of the need to extend thesaurus content.[54] An important component of data management,

[50] In June 2017, *HTE3* was updated to include the means to generate heatmaps and sparklines, visualizations that indicate degrees of lexicalization throughout the recorded history of the English language.

[51] The nine co-ordinate categories for peace are the category "13.01 Peace, state of law and order" and the subcategories "13|01 Peace, absence of dissension", "13|02 Unanimity, concord, agreement", "13|03 n. Good terms, rapport", "13|04 Peace, freedom from hate", "13|05 Peace among dwellings", "13|06 Accordant, not at variance", "13|07 Peacefully, peaceably", and "13|08 To agree, settle".

[52] Kay, 'Food as a Fruitful Source of Metaphor', p. 77.

[53] See Appendix 2.A.

[54] This functionality has been requested and used by researchers in the project 'Exploring Early Medieval English Eloquence' (see Appendix 2.A). For the need to extend thesaurus content, see section 2.3.

necessary to support the extension of data by researchers, concerns personal additions, such as annotations and labelling. Researchers who supplement a historical language thesaurus with such additional knowledge should retain full control over their own data, including the ability to display or hide their annotations, create backups of their additions, and share their data with others.

A second component of data management, applicable to historical language thesauri extended by other data sources, is the means to select which of these sources are deemed relevant for researchers' explorations.[55] Web-based editions of thesauri should allow for sets of information to be combined, based on the user's selection, for viewing and analyses, thus supporting researchers in their onomasiological explorations of lexis and any salient features captured. An advanced form of this functionality ought to support choosing for a certain revision of a data source, too, in cases where multiple are available. Revisions of digital historical language thesauri are certainly not shunned by their editors.

At its initial publication, the first digital edition of *TOE* contained some corrections to its printed counterpart. These included changes of usage information and additions of new words, on the basis of new knowledge, which stemmed "largely from completed sections of the Toronto Dictionary of Old English".[56] The exact alterations were left unspecified, though these will have mostly concerned lexical items starting with the letter F: between the publication of the second impression in print in 2000 and the first publication of the electronic edition in 2005, the Toronto dictionary only published its findings on Old English words starting with that letter.[57] As for any further updates to electronic editions, Kay mentioned that the taxonomy "will, of course, be subject to rolling revision" as the electronic environment allows for such changes.[58] Indeed, the number of items the online *TOE* contains has increased from the 50,706 items Kay mentions in her article for the initial release of the electronic edition to 51,483 items on 26 May 2017.[59]

When digital thesauri are subject to ongoing revisions, it may present difficulties for scholars to work with them.[60] Firstly, it could mean that the contents change whilst the scholar is doing research that involves the contents of the thesaurus. A lexical item could no longer be included under the category it used to be present in, and a category might have moved within the topical system. In effect, such changes during research entails scholars work with material that is ever in progress and for which it becomes difficult to discuss that reference body

[55] This functionality has been requested and used by researchers in the project 'Exploring Early Medieval English Eloquence' (see Appendix 2.A).

[56] Kay, '*A Thesaurus of Old English* Online', pp. 36–40.

[57] The section on the letter F was published in 2004. See the 'Publications' section of *DOE*. Accessed on October 31, 2016.

[58] Kay, '*A Thesaurus of Old English* Online'.

[59] An export of the *TOE* database as it was on 26 May 2017 has kindly been provided to the author by the University of Glasgow.

[60] See also Allan's remark on working with information from *HTE*: "It seemed preferable, and more theoretically justifiable, to work with the data as it existed at a particular stage of *HTE*, whilst acknowledging that this may be incomplete. This is especially the case given the current revision of the *OED*, which will in turn affect *HTE* data and may lead to a number of insertions and changes in later editions" (*Metaphor and Metonymy*, p. 20).

as a whole or even partially. Secondly, such revisions can make it more difficult to ensure verifiability of performed research. One of the important reasons why articles include references to the material they employed is that this practice allows others to review their work, verify whether the source material is accurate and reliable, and that the results are justifiable and could be reproduced if so desired.[61] With a reference body that is unstable, references to its contents may not provide the verifiability desired. Note that for publications in print, no such difficulties are present. Scholars can select an edition available at the moment of writing, employ that particular edition consistently, and have their references include the information on which edition should be consulted for purposes of verification.

In order to ensure that scholars can choose a particular edition or revision of a historical language thesaurus, the publication environment will need to support a form of versioning. That is to say, the thesaurus should be accessible under a system that allows users to view the taxonomy in each of its stages, enabling scholars to view and refer to the taxonomy in the state it was for a certain publication that made use thereof. Some digital reference bodies already provide such means to look into earlier states. Entries of the third edition of the *OED*, for instance, contain links to the corresponding entries from the previous edition.[62] Maintaining separate editions online, similar to how this is the case for publications in print, will also open the possibility to pinpoint which changes have been made from one revision to the next, where they are located in the topical system, and possibly even for what reasons these have been made. Such information could help scholars in determining how up-to-date the taxonomy is, and whether they agree with the adjustments made, rather than having to analyse such matters themselves. Presently, such functionality is missing from all historical language thesauri in an electronic form treated here.

2.6. Conclusion

Drawing on existing historical language thesaurus editions, academic reviews, and research employing these resources, the current chapter has addressed functionality desired by researchers. An overview of the five functionalities that have been discussed is presented in Table 2.1. Inclusion of this functionality varies in the existing editions of the historical language thesauri analysed. Whereas all editions offer navigation (through the topical system and through an index) and resource views, the ability to extend their content is found in only one thesaurus (i.e., *HTS*) and is limited to sharing user comments. The effect of this lack of options to extend these thesauri is that, unsurprisingly, the need and mechanisms for data management are in these cases absent, too. Similarly, analysing thesaurus content through automated means is, across all thesauri, either minimal or vastly underexplored when considering the additional

[61] See, for instance, the common ethics and responsibilities within science regarding the treatment of data and the sharing of research results as discussed in Gauch Jr, *Scientific Method in Brief*, pp. 226–30.

[62] See *OED Online*.

Code	Name	Description
R1	Navigation	Approaching a thesaurus should be possible in two manners: through its overarching taxonomy or through the lexis it organises.
R2	Resource views	Complete overviews of available information are to be presented on any given resource within a thesaurus that a user chooses to inspect.
R3	Extension	Thesaurus content should be extendable, allowing users to connect additional information to existing content.
R4	Analyses	Statistical analyses, utilizing the onomasiological structure of the thesaurus and features of the lexis it contains, should be made possible.
R5	Data management	Users must have full control over their own data and the ability to select which data sources are deemed relevant for their explorations, allowing sets of information to be combined for viewing and analysis.

Table 2.1.: Functionality required by researchers of historical language thesauri.

information that the semantic frameworks formed by these lexicographic works have the potential to yield. In short, the functionality identified here as required for research and education – navigation, resource views, extension, analyses, data management – should open new research avenues when incorporated into the dissemination of historical language thesauri. Sharing this set of functionality with researchers will greatly facilitate exploring the lexis of a historical language through a semantic lens.

References

Allan, K., *Metaphor and Metonymy: A Diachronic Approach*, Publications of the Philological Society 42 (Oxford, 2008).

Århammar, N., 'A Frisian Supplement to Buck's Dictionary of Indo-European Synonyms?', *Us Wurk* 21-22 (1972-1973), 241-5.

Atkins, B. T. S. and M. Rundell, *The Oxford Guide to Practical Lexicography* (Oxford, 2008).

Bremmer Jr, R. H., 'Treasure Digging in the Old English Lexicon', Review of *TOE2*, *NOWELE* 40 (2002), 109–14.

Brewer, C., 'Labelling and Metalanguage', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 488–500.

BTH = The Bilingual Thesaurus of Everyday Life in Medieval England, eds. (Glasgow, 2019). https://thesaurus.ac.uk/bth/. Accessed: 15 March 2022.

Busse, B., 'A Celebration of Words and Ideas: The Stylistic Potential of the *Historical Thesaurus of the Oxford English Dictionary*', Review of *HTE1* and *HTE2*, *Language and Literature* 21.1 (2012), 84–92.

Cavill, P., 'Names and Things in Anglo-Saxon and Early Norman England', Review of *TOE1* and *Words, Names and History: Selected Writings of Cecily Clark* ed. by Peter Jackson, Nottingham Medieval Studies 41 (1997), 186–91.

Conner, P. W., Review of *TOE1*, *Speculum* 73.3 (1998), 887–9.

Cruse, D. A., *Lexical Semantics* (Cambridge, 1986).

Crystal, D., *Words in Time and Place: Exploring Language through* The Historical Thesaurus of the Oxford English Dictionary (Oxford, 2014).

Dance, R., Review of *TOE1*, *Medium Ævum* 66.2 (1997), 312–13.

Diller, H.-J., 'Measuring the Growth of Semantic Fields: The Case of the English Emotion Lexicon', in *Cognition in Language: Volume in Honour of Professor Elżbieta Tabakowska*, eds. W. Chłopicki et al.(Kraków, 2017), pp. 574–96.

DOE = *Dictionary of Old English: A to I Online*, eds. A. Cameron et al. (Toronto, 2018). https://tapor.library.utoronto.ca/doe/.

DSSPIEL = Buck, C. D., *A Dictionary of Selected Synonyms of the Principal Indo-European Languages: A Contribution to the History of Ideas* (Chicago, 1949).

Gauch Jr, H. G., *Scientific Method in Brief* (Cambridge, 2012).

Görlach, M., Review of *TOE1*, *Anglia* 116.3 (1998), 398–401.

Hartmann, R. R. K. and G. James, *Dictionary of Lexicography* (London, 1998).

HTE = *Historical Thesaurus of English*; see *HTE1*, *HTE2*, *HTE3*.

HTE1 = *Historical Thesaurus of the Oxford English Dictionary*, 2 vols., eds. C. Kay et al. (Oxford, 2009).

HTE2 = *Historical Thesaurus of the Oxford English Dictionary*, eds. C. Kay et al. (2010). http://oed.com/thesaurus.

HTE3 = *The Historical Thesaurus of English*, eds. C. Kay et al. (Glasgow, 2014). http://historicalthesaurus.arts.gla.ac.uk.

HTS = *Historical Thesaurus of Scots*, ed. S. Rennie (Glasgow, 2015). http://scotsthesaurus.org.

Ilson, R. F., 'On the Historical Thesaurus of the Oxford English Dictionary', Review of *HTE1*, *International Journal of Lexicography* 24.2 (2011), 241–60.

Kay, C., Review of *ShT*, *International Journal of Lexicography* 9.1 (1996), 71–6.

Kay, C., '*A Thesaurus of Old English* Online', *Old English Newsletter* 38.3 (2005), 36–40.

Kay, C., 'Food as a Fruitful Source of Metaphor', in *Mapping English Metaphor Through Time*, eds. W. Anderson et al. (Oxford, 2016), pp. 66–78.

Kay, C. and M. Alexander, 'Diachronic and Synchronic Thesauruses', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 367–80.

Learning with the Online Thesaurus of Old English, eds. C. Hough and C. Kay (Glasgow, 2007). http://oldenglishteaching.arts.gla.ac.uk. Accessed: 31 October 2016.

LSM = Coleman, J., *Love, Sex and Marriage: A Historical Thesaurus*, Costerus New Series 118 (Amsterdam, 1999).

Mapping English Metaphor Through Time, eds. W. Anderson et al. (Oxford, 2016).

McCracken, J., 'The Exploitation of Dictionary Data and Metadata', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 501–14.

Momma, H., Review of *TOE2*, *Notes and Queries* 50.1 (2003), 79–80.

Norri, J., 'Regional Labels in Some British and American Dictionaries', *International Journal of Lexicography* 9.1 (1996), 1–29.

Mapping Metaphor with the Historical Thesaurus (Glasgow, 2015). http://mappingmetaphor.arts.gla.ac.uk. Accessed: 31 October 2016.

OED = *Oxford English Dictionary*; see *OED1*, *OED2*, *OED3*, *OED Online*.

OED1 = *The Oxford English Dictionary*, eds. J. A. H. Murray et al. (Oxford, 1884–1933).

OED2 = *The Oxford English Dictionary*, eds. J. A. Simpson and E. S. C. Weiner, 2nd edn (Oxford, 1989).

OED3 = *The Oxford English Dictionary*, ed. J. A. Simpson. http://oed.com.

OED Online = *Oxford English Dictionary Online*. http://oed.com. Accessed October 18, 2016.

Porck, M. H., 'Growing Old among the Anglo-Saxons: The Cultural Conceptualisation of Old Age in Early Medieval England' (Ph.D dissertation, Leiden University, 2016). https://hdl.handle.net/1887/39136.

The Scots Thesaurus, eds. I. Macleod et al. (Aberdeen, 1990).

ScT = *The Scots Thesaurus*, eds. I. Macleod et al. (Aberdeen, 1990).

ShT = Spevack, M., *A Shakespeare Thesaurus* (Hildesheim, 1993).

Stolk, S., 'Welcoming the *Thesaurus of Old English Statistics*: *The Thesaurus of Old English* and the Vocabulary of Greetings' (M.Phil. dissertation, Leiden University, 2013). https://hdl.handle.net/1887/31760.

Svensén, B., *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making* (Cambridge, 2009).

TOE = *A Thesaurus of Old English*; see *TOE1*, *TOE2*, *TOE3*, *TOE4*.

TOE1 = *A Thesaurus of Old English*, eds. J. Roberts et al., 2 vols., King's College London Medieval Studies XI (London, 1995).

TOE2 = *A Thesaurus of Old English*, eds. J. Roberts et al., 2 vols., Costerus New Series 131, 2nd edn (Amsterdam, 2000).

TOE3 = *A Thesaurus of Old English*, eds. J. Roberts et al. (Glasgow, 2005). http://libra.englang.arts.gla.ac.uk. Accessed February 1, 2013.

TOE4 = *A Thesaurus of Old English*, eds. J. Roberts et al. (Glasgow, 2015). http://oldenglishthesaurus.arts.gla.ac.uk. Accessed March 15, 2022.

Wierzbicka, A., *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese*, Oxford Studies in Anthropological Linguistics 8 (New York, 1997).

Appendix 2.A: Functionality elicited in 'Exploring Early Medieval English Eloquence'

This appendix offers an overview of functionality required of historical language thesauri for research. The functionalities listed here have been elicited in the project 'Exploring Early Medieval English Eloquence', which utilizes *A Thesaurus of Old English*, through dedicated stakeholder meetings, workshops, and feedback based on preliminary results in research and education. This project and its case studies are discussed in Chapter 8. Functionality marked as novel in the table below is, at the time of writing, unavailable in any of the thesaurus editions analysed in Chapter 1 and Chapter 2. For each functionality request, a footnote describes whether and when it was implemented in the web application Evoke (described in Chapter 7) and in which of the case studies (discussed in Chapter 8) this functionality was used.

Functionality	Novel
Navigation (R1)	
Navigation via the thesaurus structure: The user must be able to navigate the topical structure of the thesaurus and open a given category in a resource view.[63]	-
Navigation via the thesaurus index: The user must be able to search for specific elements (i.e., lexical items, categories, labels) and open a given one in a resource view.[64]	-
Show subthesaurus: Show only those categories and items of interest, based on a selection of features made by the user. (This requirement holds for resource views as well as navigation.)[65]	+
Advanced search: Restrict searches to items based on a selection of features made by the user.[66]	-
Resource views (R2)	
For any resource, offer basic information: Offer rudimentary information on a resource, such as its name, identifier (i.e., metadata), and position within the topical structure.[67]	-

[63] This functionality has been implemented in Evoke (since v1.0.0) and used in the following case studies described in Chapter 8: Dekker (education); Depuydt and De Does; Fletcher; Khan et al.; Porck; Porck and Stolk (education); Van Baalen; Van de Poel and Stolk.

[64] Implemented in Evoke (since v1.0.0) and used in: Dekker (education); Depuydt and De Does; Khan et al.; Porck; Porck and Stolk (education); Van Baalen; Van de Poel and Stolk.

[65] Not yet implemented in Evoke (v1.4.1). This functionality has been requested by four researchers participating in the EEMEE workshops.

[66] Not yet implemented in Evoke (v1.4.1). This functionality has been requested by one researcher participating in the EEMEE workshops.

[67] Implemented in Evoke (since v1.0.0) and used in: Dekker (education); Depuydt and De Does; Fletcher; Khan et al.; Porck; Porck and Stolk (education); Van Baalen; Van de Poel and Stolk.

Functionality	Novel
For a category, show lexical items allocated to it: Show, for a category, which lexical senses are positioned at this thesaurus category.[68]	-
For a lexical item, show senses related through polysemy: Show, for a lexical sense and for its lexeme, all senses attributed to the lexeme in question.[69]	+
For a lexical sense, show synonyms: Show, for a lexical sense, which synonyms are available in the language. If multiple languages are available, show lexical senses that lexicalize the same concept (i.e., are translations).[70]	-
For a lexical item, show its labels: Show labels in the resource view, both original and ones newly introduced by the user, assigned to lexical items. (This requirement applies to the index as well as to resource views.)[71]	-
For a lexical item, show its language: Show the language in the resource view, besides part of speech, attributed to lexical items. (This requirement applies to the index as well as to resource views.)[72]	-
For a label, show resources marked by it: Show, for a label, which lexical items or other resources were marked with this label.[73]	+
For a category, show associations: Show, for a category, which categories are evoked by other senses of the lexical items found at this thesaurus category (i.e., senses related through polysemy).[74]	+
Extension (R3)	
Annotate and label content: Users should be able to annotate lexical items (and other elements) with custom labels.[75]	+
Link bodies of knowledge: Users should be able to link data from another work to data contained in a historical language thesaurus.[76]	+

[68] Implemented in Evoke (since v1.0.0) and used in: Dekker (education); Depuydt and De Does; Fletcher; Khan et al.; Porck; Porck and Stolk (education); Van Baalen; Van de Poel and Stolk.

[69] Implemented in Evoke (since v1.1.0) and used in: Dekker (education); Depuydt and De Does; Khan et al.; Porck; Porck and Stolk (education); Van Baalen; Van de Poel and Stolk.

[70] Implemented in Evoke (since v1.0.0) and used in: Dekker (education); Porck; Porck and Stolk (education); Van Baalen; Van de Poel and Stolk.

[71] Implemented in Evoke (since v1.2.0) and used in: Fletcher; Khan et al.; Porck; Van Baalen.

[72] Implemented in Evoke (since v1.3.0) and used in: Depuydt and De Does; Van de Poel and Stolk.

[73] Implemented in Evoke (since v1.3.0) and used in: Dekker (education); Fletcher; Khan et al.

[74] Implemented in Evoke (since v1.4.0) and used in: Dekker (education).

[75] Implemented in Evoke (since v1.3.0) and been used in: Dekker (education); Fletcher; Khan et al.; Porck; Van Baalen.

[76] Implemented in Evoke (since v1.0.0) and used in: Depuydt and De Does; Porck; Van de Poel and Stolk.

Functionality	Novel
Match word list: Match an existing list of lexical items, from one knowledge body, against ones found in the thesaurus.[77]	+
Analyses (R4)	
Provide basic analyses: Users should have access to basic onomasiological analyses on the composition of semantic fields (e.g., on the basis of part of speech or number of items found over the various subordinate categories) when viewing a category.[78]	-
Provide advanced analysis: Users should be able to perform advanced onomasiological analyses for the items of their interest, based on a selection of features made by the users themselves.[79]	+
Contrast advanced analyses: Users must be able to contrast onomasiological analyses of one set of features versus another.[80]	+
Top statistics: Provide a top 10 on categories that contain the most lexical items, possibly restricted based on a selection of features made by the user.[81]	+
Data management (R5)	
Manage own data: Users must have full control over their own data: the ability to hide or view their additions, create backups, and restore backups.[82]	+
Select data sources: Users must be able to select which data sources are deemed relevant for their explorations, allowing sets of information to be combined for viewing and analysis.[83]	+

Table 2.A.1.: Initial set of functionality elicited for research.

[77] This functionality is perhaps best facilitated through tooling specifically designed for aligning two bodies of knowledge as opposed to through a thesaurus edition and its functionality. Chapter 7 describes three such tools that have been used to link, or align, different lexicographic works: a custom alignment tool by Porck, Excel spreadsheets by Van de Poel and Stolk, and a Lex'it-based linking tool by Depuydt and De Does.

[78] Implemented in Evoke (since v1.1.0) and used in: Dekker (education); Porck and Stolk (education).

[79] Implemented in Evoke (since v1.3.0) and used in: Porck; Van Baalen; Van de Poel and Stolk.

[80] Implemented in Evoke (since v1.4.0) and used in: Porck; Van Baalen; Van de Poel and Stolk.

[81] Not yet implemented in Evoke (v1.4.1). This functionality has been requested by one researcher participating in the EEMEE workshops.

[82] Implemented in Evoke (since v1.3.0) and used in: Dekker (education); Depuydt and De Does; Fletcher; Khan et al.; Porck; Van Baalen; Van de Poel and Stolk.

[83] Implemented in Evoke (since v1.4.0) and used in: Dekker (education); Depuydt and De Does; Khan et al.; Porck; Van Baalen; Van de Poel and Stolk.

PART II

Historical Language Thesauri and a Digital Form on the Semantic Web

Chapter 3

3. A digital form for historical language thesauri on the Semantic Web

Computers store knowledge in various digital forms. Each form has its strengths and weaknesses, making some forms better suited for a specific purpose than others. The PDF format, for instance, is highly suitable for storing readable documents, whereas the ZIP format is a better fit for bundling and compressing files. Similarly, the form in which a historical language thesaurus is available can facilitate or hinder its intended use. Offering a thesaurus as a PDF document, for instance, allows users to read it, but hinders automated queries over the content. Similarly, some database technologies with which thesauri can be stored may facilitate querying but be less suited for reuse and expansion of the content. The form that historical language thesauri should be published in on the Web is therefore important to consider and is the main topic of this chapter.

The investigation presented in this chapter was performed in December 2017 and led to the selection of a digital form (constituted by a set of so-called data vocabularies) appropriate for representing historical language thesauri on the Web. The combination and use of these vocabularies has, in a wider community towards standardization of the representation of linguistic and lexicographic resources on the Web, recently been termed Linguistic Linked Data.[1] This name will be used from Chapter 6 onwards to refer to the digital form.

The outline of this chapter is as follows. Section 3.1 reflects on existing publications of historical language thesauri and their shortcomings for both computational efforts and reuse of the knowledge they contain. Best practices for publishing data on the Web are discussed to which, ideally, a data form for historical language thesauri conforms. Subsequently, section 3.2 explores the Semantic Web as a candidate technology for sharing lexicographic material on the Web. Section 3.3 defines a methodology to establish an appropriate Semantic Web form of historical language thesauri. To this end, it lists three criteria for selecting appropriate data vocabularies, sets of terminology in which Semantic Web content can be expressed. Sections 3.4–3.7 then provide analyses of key information components found in thesauri – the topical system, lexical senses, synonymy, and sense attributes – and thereby construct a combination of suitable data vocabularies in which to express such lexicographic resources.

[1] Cimiano et al., *Linguistic Linked Data: Representation, Generation and Applications*. Chapter 7, section 7 of the present thesis discusses recent efforts surrounding Linguistic Linked Data and software developed for exploring resources in this format.

3.1. Historical language thesauri and their current forms

The existing historical language thesauri of Scots and English, discussed in Chapter 1, are offered to readers in paper form or, online, as webpages on the World Wide Web. Both of these forms are geared towards a particular goal: to provide material to users in a readable way. However, for filtering, querying, and other computational processing, these forms are far from suitable. The cause of this disadvantage is, as James McCracken, Principal Editor of the *Oxford English Dictionary*, phrases it, that "human readability always seems bound to conflict with computational parseability at some point".[2] The structure of webpages, for instance, may be machine parseable, but this parseability is meant for visualization of the content in browsers as opposed to other computational processing of the actual knowledge that the webpage documents. In short, both print and webpages are lacking in supporting many of the functional needs for thesauri that were identified in the previous chapter. Webpages are not the only form, though, in which the content of online editions of historical language thesauri exists.

The electronic pages of *TOE*, *HTE*, *HTS*, and *BTH* are created dynamically by combining and arranging bits of information taken from underlying electronic databases.[3] Such practice ensures that it is relatively easy to, for example, change the visualisation of the same knowledge in future updates of the website. A change of logo, a different font, making headings bold rather than underlined — these adjustments are then quickly made. Indeed, it is for such reasons considered good practice to separate presentation from the knowledge behind it.[4] The online editions of historical language thesauri are far from the only lexicographical works that employ a database and generate a presentation on top of it. In fact, as McCracken points out, there is an "increasing concensus" that a separation of knowledge from presentation form "is how good lexicography ought to proceed" — for instance by first compiling a lexical database and afterwards transforming it into a human-readable dictionary or thesaurus (whether in print or electronic).[5]

"Draw[ing] directly on the underlying lexical database" of an electronic lexicographic resource, instead of on its presentation, facilitates computational efforts and reuse of the lexical knowledge in other digital bodies.[6] In other words, the form in which databases of thesauri are published is of importance for their utilization. Unfortunately, the databases behind *TOE*, *HTE*, *HTS*, and *BTH* are not made accessible to the public. As a consequence, executing new queries (such as how many poetic senses are positioned within a given category and all its subcategories) is not possible for users. Additionally, even if these databases were made accessible, there is still the matter of whether their format and access

[2] McCracken, 'The Exploitation of Dictionary Data and Metadata', p. 504.

[3] The digital editions of historical language thesauri maintained by the University of Glasgow, including *TOE4* and *HTE3*, employ MySQL database technology. See the section 'Creation of the *Thesaurus*' of *TOE4*.

[4] See, for instance, the relation between the HTML and CSS standards: 'HTML & CSS'.

[5] McCracken, 'The Exploitation of Dictionary Data and Metadata' p. 504.

[6] Ibid.

protocols facilitate users and enable applications to work with their content.

The digital form in which data is best made available is one of the subjects in guidelines for publishing data on the Web. The FAIR principles, for instance, state a number of requirements to make data findable, accessible, interoperable, and reusable.[7] Similarly, the World Wide Web Consortium (W3C) published their 'Data on the Web Best Practices' (DWBP), which outline how data can be shared to "facilitate interaction between publishers and consumers". Applying these guidelines should net a number of benefits, including reuse and interoperability of the published information. The core advice on digital forms in these guidelines consists of using machine-readable standardized formats for data, reusing common data terminology for the domain, and having persistent identifiers for the dataset and bits of knowledge therein. The databases behind *TOE*, *HTE*, *HTS*, and *BTH* – even if they were to be opened up for public access – do not adhere to the two lastmentioned practices, hampering reuse and interoperability of the information. These databases are based on MySQL technology,[8] which lacks innate features to represent lexical information by reusing common terminology and assigning persistent identifiers in the form of Internationalized Resource Identifiers (IRIs) per database entry (e.g., per lexical sense).[9] IRIs are assigned by the websites, instead, which form stable web addresses (or URLs) for categories – but not for lexical senses or other material – based on information captured in the databases.[10] The question remains, therefore, what data form instead might be better suited to share historical language thesauri on the Web?

3.2. Lexicography and the Semantic Web

In his discussion on the exploitation of dictionary data, James McCracken not only addresses the question of why and how one would want to access such data for further computation, but also what an appropriate digital form would be to achieve such ends. He notes: "It makes sense to adopt a single formalism—if not RDF/OWL, then certainly a model that can be readily converted to and from RDF/OWL".[11] One of the arguments in favour of this formalism is that lexicographical works, McCracken notes, have graph-like characteristics.[12] Lexical senses, for instance, can be seen as nodes of such a graph, connected through such relations as hyponymy and synonymy. By expressing these characteristics in an information graph, the utility of the captured knowledge is enhanced for machine processing.[13] That RDF/OWL is seen as the "standard formalism for information graphs" – and many linguistic

[7] Wilkinson et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'.

[8] See the section 'Creation of the *Thesaurus*' of *TOE4*.

[9] *MySQL 5.7 Reference Manual*.

[10] The exception to this statement is the website of the *HTE* edition incorporated into the *OED*, which coins web addresses that identify lexical senses as well as categories.

[11] McCracken, 'The Exploitation of Dictionary Data and Metadata', p. 512.

[12] Ibid.

[13] Ibid., p. 511.

projects have already successfully adopted it or are in the process of adopting it – merits its recommendation as digital form in which to share lexicographical resources.[14] But what exactly is RDF/OWL and what are the characteristics attributed to this "standard formalism for information graphs"?

RDF and OWL are fundamental standards within the Semantic Web.[15] This Web is, in essence, one of linked data, built on top of a set of data standards and technologies that aim to add "well-defined meaning" to information.[16] These standards and technologies provide an "infrastructure for publishing, storing, retrieving, reusing, integrating, and analyzing data".[17] Its data form, open for anyone to use, is comprised of statements (or triples) that together form a network of information, one in which concepts are identified by IRIs (often in the form of web addresses or URLs).[18] The use of IRIs allows for capturing and identifying data, reusing terminology defined elsewhere, and connecting information found in different digital resources. In effect, this identification mechanism enables thesaurus content to be reused, extended with custom labels and with links to other digital resources. These characteristics of the data form offer intrinsic support for many of the DWBP best practices and facilitate achieving the remaining ones. The underlying data format for Semantic Web information, RDF, is therefore mentioned explicitly in the DWBP documentation as highly suitable for publications on the Web. OWL provides an additional layer of expression that can be used on top of RDF for situations in which highly formal definitions and inferencing mechanisms are required.[19] For more informal levels of expression, other Semantic Web standards, such as SKOS, are an alternative to OWL.[20]

Together, Semantic Web standards offer a well-defined formalism for information graphs. Moreover, they are said to support the functionality desired of historical language thesauri as identified in Chapter 2, since the infrastructure of the Semantic Web allows data to be "retrieved, accessed, reused, and integrated in a meaningful way".[21] Perceived benefits in using this data form are mentioned by Christian Chiarcos et al.[22] One of these benefits is the ability to merge different datasets, or relate different perspectives and conceptualizations of similar data, in order to obtain a combined set of data that is validly formatted.

---

[14] Ibid.

[15] See 'Semantic Web'. For an accessible introduction to the Semantic Web standards and uses, see Allemang and Hendler, *Semantic Web for the Working Ontologist*.

[16] Semantic Web technology was intended to add "well-defined meaning" to information on the Web (Berners-Lee et al., 'The Semantic Web'). Examples of such meaning are relations of hyponomy and of incompatibility. The former can be expressed through the subclassification in the RDFS vocabulary ('RDF Schema 1.1'); the latter through disjointness of classes in the OWL vocabulary ('OWL 2 Web Ontology Language').

[17] Janowicz et al., 'Why the Data Train Needs Semantic Rails', p. 5.

[18] 'RDF 1.1 Concepts and Abstract Syntax'.

[19] 'OWL 2 Web Ontology Language: Document Overview'.

[20] 'SKOS Simple Knowledge Organization System Reference'. This chapter will explore which standards are appropriate for use in a Semantic Web form of historical language thesauri from section 3.3 onwards.

[21] Janowicz et al., 'Why the Data Train Needs Semantic Rails', p. 13.

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

Thus, thesauri and sets of data elaborating on them can be queried in unison.[23] A second benefit is an increased level of interoperability. Using standardized terminology in describing linguistic data increases a shared understanding of that data and facilitates their interpretation by software. Moreover, the use of IRIs as identifiers ensures data can be linked without the need for duplication of information from one set into another. The ability to link (or reference) in such a manner is valuable for historical language thesauri, too, since some of these resources are subject to licenses intended for viewing only, stipulating that users are not allowed to copy or download a substantial portion of their content (e.g., *TOE* and *HTE*). By adopting IRIs in published thesauri, their users should be able to explore and extend these resources, engaging with the content offered, without infringing on such licenses. In short, this data form is promising for representing historical language thesauri on the Web. However, since such a thesaurus is yet to be captured in this form, the question remains as to how a Semantic Web form of these lexicographic resources should be obtained.

3.3. Obtaining a Semantic Web form

Considering the existing digital forms of historical language thesauri do not yet include a Semantic Web form, it could be argued that such a form could be modelled in any way one would see fit. However, as DWBP best practice 15 and FAIR principle I2 argue, reuse of common terminology in expressing knowledge increases data interoperability and chances at reuse.[24] Such terminology for expressing the semantics of digital information is found in data vocabularies (which in the Semantic Web and Linked Data communities are sometimes referred to as metamodels, metavocabularies or, simply, vocabularies).[25] DWBP indicates that one can locate appropriate data vocabularies for the Semantic Web through repositories, such as the *Linked Open Vocabularies* (*LOV*) repository.[26] The best practices put forward for Semantic Web data specifically, the 'Best Practices for Publishing Linked Data', also suggest using directories, such as *LOV*, or, alternatively, to look into already published datasets. It is apparent why both documents recommend *LOV*. The repository contains a useful overview of existing Semantic Web data vocabularies, their terminology, and locations of access.[27] Moreover, *LOV* offers functionality to search for individual terms, query over the data vocabularies, and visualise them. At the time of writing, the

[23] This capability facilitates extending original thesaurus content (see requirement R3 in Chapter 2).

[24] See 'Data on the Web Best Practices' and Wilkinson et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'.

[25] DWBP, for instance, employs both *data vocabulary* and *vocabulary*. Articles by Van Assem et al. adopt the term *metamodel* instead. See Van Assem et al., 'A Method for Converting Thesauri to RDF/OWL'; and Van Assem et al., 'A Method to Convert Thesauri to SKOS'.

[26] 'Data on the Web Best Practices'.

[27] *Linked Open Vocabularies*. For an introduction to the website and its usefulness, see the following article: Vandenbussche, 'Linked Open Vocabularies'.

repository lists well over 500 data vocabularies.[28] Knowing where to find existing data vocabularies, one should address the question of which criteria ought to be used in selecting vocabularies for bringing historical language thesauri to the Semantic Web.

Two crucial criteria for selecting a data vocabulary are named by Mark van Assem et al., who have converted many digital resources to Semantic Web forms successfully.[29] Firstly, the chosen terminology should facilitate preserving all knowledge and intended semantics. After all, loss of meaning or misrepresentations of the content may hinder correct interpretations of the content and its reuse. Secondly, the terminology should be a standard (or extend a standard) in order to promote interoperability with other resources similar in nature. In addition to these criteria from Van Assem et al., the present chapter takes a third into account: coherency between terms employed should be well understood. Mixing and matching individual terms from a plethora of data vocabularies, for instance, is likely to net ambiguous or even unknown connections between individual terms. Datasets can be interpreted better when terminology from a few select vocabularies is used instead. In such cases, their cohesion is known and well understood. In short, three criteria are considerd key in selecting appropriate data vocabularies: coverage, standardization, and coherency. Using these three criteria, the next sections will discuss appropriate data vocabularies for each component of historical language thesauri identified in Chapter 1: their topical system, lexical senses, and the relation of synonymy between senses. Sense attributes commonly found in these thesauri (i.e., part of speech, usage features, and language) are covered after these main components.

3.4. Semantic Web form for topical systems

On the Semantic Web, the term "topical system" is not as commonplace as the phenomenon itself. To illustrate, the term *topical* occurs in only four *LOV* vocabularies — not one of which has a clear focus on capturing topical systems.[30] This result begs the question of what terminology is employed on the Semantic Web instead. The answer can be found in one of the meanings of the word *thesaurus*. As Reinhard Hartmann states, one sense of the word is that of a 'terminological database' or 'index'.[31] The *OED* defines this particular sense, first attested in 1957, as follows: "A classified list of terms, esp. key-words, in a particular field, for use in indexing and information retrieval."[32] In essence, this sense of *thesaurus* coincides with what is known as the topical system in historical language thesauri: concepts or labels that are arranged in a hierarchical manner, typically based on the semantics of these concepts, in order to index

[28] This chapter was written in December 2017. The *LOV* repository has grown since and, on 18 April 2022, contains information on 774 vocabularies.

[29] See Van Assem et al., 'A Method for Converting Thesauri to RDF/OWL'; and Van Assem et al., 'A Method to Convert Thesauri to SKOS'.

[30] *LOV*, s.v. 'dbpedia-owl', 'gold', 'lmm1', 'umbel'.

[31] *Encyclopedia of Language & Linguistics*, s.v. 'Thesauruses', by Hartmann.

[32] *OED*, s.v. 'thesaurus, n.', sense 2c.

– or categorize – information in various forms. I will henceforth use the term *indexing thesaurus* to denote this specific sense of *thesaurus*.

Before discussing the results of searching *LOV* for appropriate data vocabularies for indexing thesauri, it should be noted that DWBP already explicitly mentions two data models that are used to capture and exchange such thesauri. These data models, which are considered relatively straightforward since "complex formalisms are most often not needed" for indexing thesauri, are the ISO 25964 data model and W3C's Simple Knowledge Organization System (SKOS) vocabulary.[33] The first-mentioned model is part of the ISO 25964 standard by the International Organization for Standardization. This standard contains guidelines on the development of indexing thesauri and proposes a data model to encourage exchange and interoperability. The current body maintaining the ISO 25964 standard, NISO, recognizes similarities between the international standard (which is divided over two parts) and SKOS.

> ISO 25964-1 essentially advises on the selection and fitting together of concepts, terms and relationships to make a good thesaurus. SKOS addresses the next step, with recommendations on porting the resultant thesauri (or other 'simple Knowledge Organization Systems') to the Web. ISO 25964-2 recommends the sort of mappings that can be established between one KOS and another; SKOS presents a way of expressing these when published to the Web.[34]

The aforementioned sources thus convey that SKOS is considered an appropriate data vocabulary to express indexing thesauri on the Semantic Web. In fact, a data vocabulary has been created specifically to supplement SKOS with terms from ISO 25964 that are not already covered by SKOS, effectively porting the data model of the ISO standard to a Semantic Web context.[35]

Next to SKOS and its ISO 25964 supplement, further data vocabularies exist that treat indexing thesauri. A search for 'thesaurus' in *LOV* yielded a number that merely refer to indexing thesauri rather than expressing them (often recommending their use)[36] or provide a definition for such reference bodies as a whole, lacking terminology to represent the actual content within.[37] Data vocabularies amongst the results that can be used to represent historical language thesaurus content are the Metadata Authority Description Schema (MADS), the Ontopic Ontology (Ontopic), the UNESKOS Vocabulary (UNESKOS), the ISO 25964 SKOS extension (ISO-THES), and, predictably, SKOS. Figure 3.1 depicts these Semantic Web vocabularies, which can be labelled candidate vocabularies for representing a topical system, and the relationship between them.

When considering the candidate vocabularies depicted in Figure 3.1, it is striking that the vast majority of them extends SKOS: these vocabularies complement SKOS with new terminology or specialize terms that already exist in

[33] 'Data on the Web Best Practices'.
[34] 'ISO 25964 Thesaurus Schemas'.
[35] Isaac and De Smedt, *ISO-THES*.
[36] *LOV*, s.v. 'ptop', 'dce', 'dcterms', 'lom', 'edm', 'gndo', 'crm', 'ecrm', 'mtlo'.
[37] These include *LOV*, s.v. 'fabio', 'iol', 'lingvo', 'crm'.

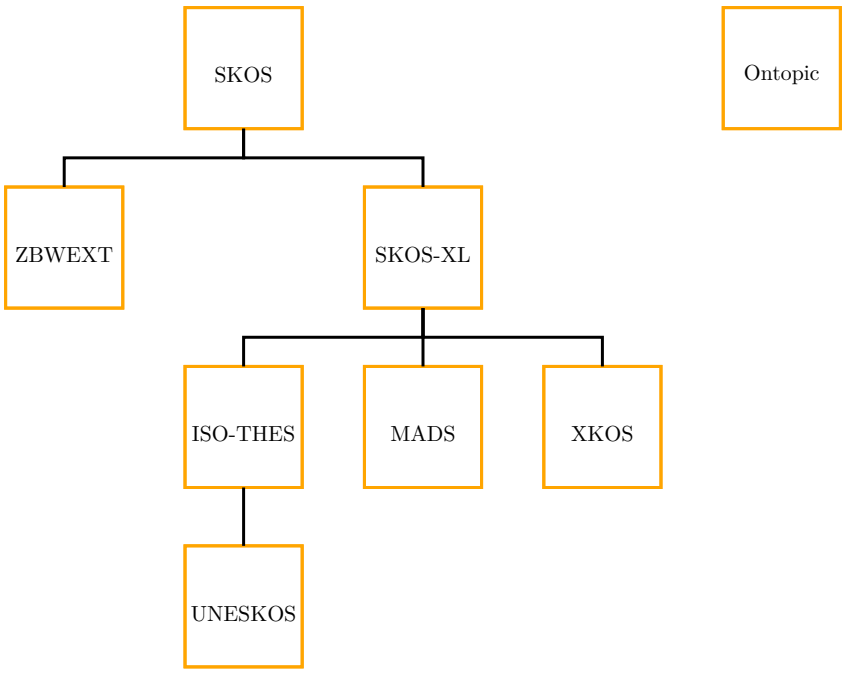


Figure 3.1.: Semantic Web vocabularies that are available for indexing thesauri (depicted as squares). Vocabularies extend higher positioned ones to which they are connected by an upward path.

SKOS. The ISO-THES data vocabulary, for instance, distinguishes three types of hierarchical relations and coins separate terms for these relations, relating each of them to the generic terminology for hierarchical relations available in SKOS.[38] The Ontopic vocabulary is an exception, posited as an alternative to SKOS for indexing thesauri. However, the maintainers of *LOV* remark that, unlike SKOS, Ontopic has seen "no visible use" on the Semantic Web.[39] The prevalence of SKOS and the lack of adoption of Ontopic, which thereby fails to meet the standardization criterion, warrants a closer inspection of SKOS as a vocabulary for expressing topical systems of historical language thesauri.

3.4.1. SKOS for topical systems

The SKOS data vocabulary is meant to express knowledge organization systems. Examples of such systems, according to its specification, are "thesauri, taxonomies, classification schemes and subject heading systems".[40] These systems have a number of traits in common — traits that can be expressed in SKOS in order to bring such knowledge to the Semantic Web environment.

The terminology in SKOS revolves around the notion of concepts. The definition of the term `Concept` is "a unit of thought", a rather general notion

[38] See the properties `iso-thes:broaderGeneric`, `iso-thes:broaderInstantial`, `iso-thes:broaderPartitive`, which are asserted to be subproperties of `skos:broader`.
[39] *LOV*, s.v. 'ontopic'.
[40] 'SKOS Simple Knowledge Organization System Reference'.

that ensures wide support for informal organizing systems.[41] Such systems all express and organize items, and SKOS therefore includes terminology to organize concepts into informal hierarchies (using the relations `broader` and `narrower` between concepts) and to create cross-references between concepts (using the relation `related`). The concepts themselves can be described using labels and notes, and aggregated into collections (`Collection`) and schemes (`ConceptScheme`).

The terminology available in SKOS is expressive enough to capture the majority of the information found in the topical system of historical language thesauri. Each category can be represented as a concept in SKOS, identified by its own IRI, labelled with a name, and placed in an informal hierarchy. The identification of a category, which encodes the position of a category within the topical system, can be expressed using the SKOS `notation` property. These identifications are, in historical language thesauri, sufficient to deduce the order of co-ordinate categories when presenting them. Editorial commentaries, too, can be expressed in SKOS through its system of notes. In short, the coverage of this data vocabulary is extensive and warrants an assessment to determine whether SKOS is sufficiently standardized, too, for representing historical language thesauri.

SKOS was finalized and published in 2009 as a recommended standard by W3C, the consortium that initiated and maintains the technological specifications for the Semantic Web. As such, SKOS is backed by an authoritative body in the Semantic Web community. The quality of the vocabulary is perhaps best illustrated through its use and reuse. As Figure 3.1 shows, SKOS has certainly seen reuse and specialization in other vocabularies, including ISO-THES and XKOS. In fact, *LOV* shows 214 data vocabularies employing SKOS, in one way or another, at the time of writing.[42]

Not just data vocabularies employ terminology from SKOS. A large number of indexing thesauri published on the Semantic Web, too, make use of this data vocabulary. A nonexhaustive list includes EuroVoc (the European Union's multilingual thesaurus), the NASA Thesaurus, the UNESCO Thesaurus, the Getty Vocabularies (including the Art & Architecture Thesaurus and the Getty Thesaurus of Geographic Names), AGROVOC (the United Nation's agricultural thesaurus), the Integrated Public Sector Vocabulary, and the Medical Subject Headings. Van Assem, who has ported several indexing thesauri to the Semantic Web (including the last two mentioned), advocates the use of SKOS in particular over coining new terminology that is completely unrelated to that found in SKOS.[43] Next to having good coverage for historical language thesauri, then, the SKOS data vocabulary is also a standard that has been adopted widely on the Semantic Web.

Although SKOS can express categories, it is not evident as to how category *types* should be captured in this data vocabulary. As mentioned in Chapter 1, the historical language thesauri *TOE*, *HTE*, and *LSM* distinguish such different

[41] Ibid.
[42] *LOV*, s.v. 'skos'.
[43] Van Assem, *Converting and Integrating Vocabularies for the Semantic Web*, pp. 145, 150.

types.[44] *TOE* distinguishes two, *HTE* three, and *LSM* four. Separation of what editors perceived as the macrostructure and the microstructure of the thesaurus, for example, are reflected in the category types.[45] The lack of clarity on capturing these distinctions in SKOS has been recognized by the authors of XKOS. The XKOS data vocabulary extends SKOS in order to express so-called classification levels.

3.4.2. XKOS for classification levels

The XKOS data vocabulary is an extension of SKOS, specifically geared towards meeting the needs of the statistical community for knowledge organization systems.[46] The specific requirements of this community include increased specificity for both hierarchies (distinguishing partonymy from hyponymy) and associations (distinguishing causal, sequential, and temporal relations). Additionally, the statistical community recognises levels in their hierarchical structures. The XKOS specification states that "levels are used as a means to identify concepts within a classification [that are] used to classify instances at the same specificity" — similar to the purpose of category types in historical language thesauri.[47]

Although the purpose of the classification levels in XKOS matches that of category types in historical language thesauri, there is a difference between the two notions. This difference makes XKOS levels unfit for expressing category types. Levels in XKOS "correspond to all those concepts that are same distance from the top of the hierarchy".[48] All top categories belong to the first level, all categories directly subordinate belong to the second level, and so on. This approach is valuable for those thesauri in which categories at a given depth of the taxonomy all share the *type* of category. However, this condition does not always hold for historical language thesauri. As demonstrated in Chapter 1, *TOE*, *HTE*, and *LSM*, distinguish different category types. Some of their categories have subordinates of different category types. In other words, an equal depth in the tree for categories does not imply the same level in the hierarchy of category types. The three thesauri use these types in a manner more flexible than XKOS is able to express: conceptual levels, which reflect the hierarchy of category types, rather than tree levels. This distinction is portrayed in Figure 3.2. Here, although the category "Permission" is four levels deep according to the definition used in XKOS, it is only a single conceptual level deep according to *HTE*. *HTE* categories at this conceptual level in are referred to as, simply, 'Categories'. The category "Disobedience", which is is visualized at the same tree level as "Permission", is nonetheless located in the second conceptual level of *HTE* rather than the first. *HTE* categories at this second conceptual level, which are thesaurus microstructure rather than its macrostructure, are referred to as 'Subcategories'.

[44] See section 1.3.3, 'Identification of categories'.
[45] See section 1.2, 'Main components'.
[46] 'XKOS'.
[47] Ibid.
[48] Ibid.

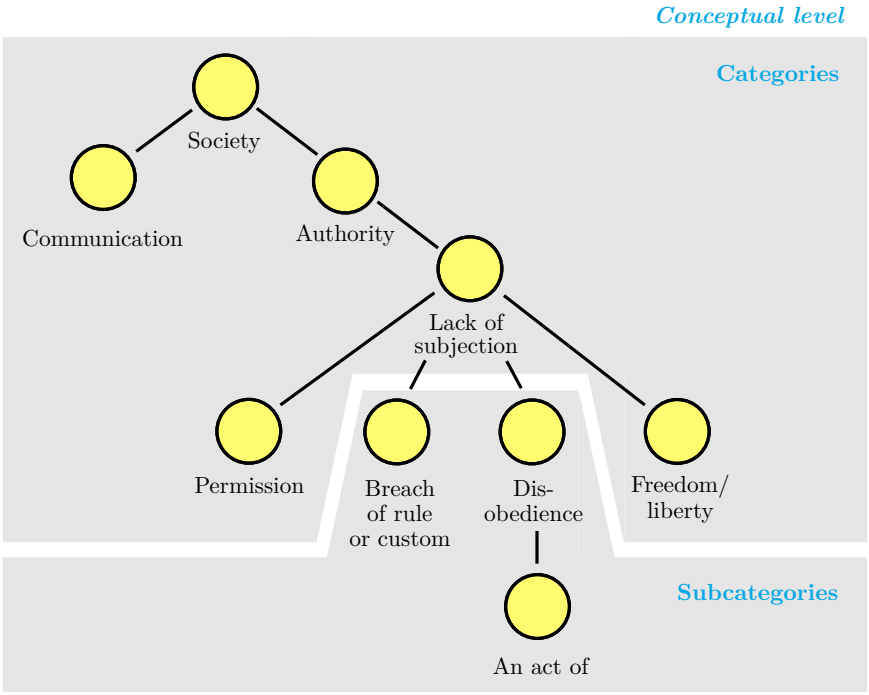


Figure 3.2.: Two conceptual levels in *HTE*.

Although XKOS levels are not appropriate for capturing category types, the manner in which the vocabulary expresses levels offers a valuable insight into how category types can be expressed. Classification levels in XKOS are specializations of concept collections in SKOS. In other words, all concepts that belong to a specific level are asserted to be members of a collection, one that XKOS calls a classification level. A classification level has a certain `depth` (from the top of the hierarchy), typically a name (such as "sections", "subsections", or "divisions"), and possibly a `notationPattern` providing details on the pattern used for the identification of concepts at this level. The concept scheme that contains the concepts may assert the number of levels it contains (using the `numberOfLevels` relation), which these are, and how they are ordered (referring to an ordered list of classification levels through the `levels` relation).

The following can be said on translating the XKOS modelling approach to the context of conceptual levels related to category types found in the historical language thesauri analysed. Firstly, much like classification levels in XKOS, conceptual levels may be posited as subtypes of SKOS collections. Each conceptual level, much like a classification level, typically has a name and may have a notation pattern. The concept schemes that contain conceptual levels would require relations similar to those found in XKOS: a relation to express the number of conceptual levels, and one that lists all available levels in their order of specificity. Such terminology, required for some of the historical language thesauri, are not yet found in the available Semantic Web data vocabularies and

may therefore have to be coined in a new one.[49]

3.5. Semantic Web form for lexical senses

In order to obtain a Semantic Web form for lexical senses, *LOV* has been searched once more for suitable data vocabularies. A search on the keyword 'sense' yields two data vocabularies in the repository that contain terminology for expressing lexical senses: the Lexicon Model for Ontologies (LEMON) and the OLiA Annotation Model for Uby Parts of Speech (UBY). The latter indicates that its term coined for lexical senses (`Sense`) is superseded by that in LEMON (`LexicalSense`), along with ten other terms, such as that for lexicon and lexical entry.[50] For this reason, this section focuses on the LEMON vocabulary and the extent to which it is suitable for capturing senses found in historical language thesauri.

3.5.1. LEMON for lexical senses

The Lexicon Model for Ontologies vocabulary (LEMON) has been designed to capture lexicons and to add their lexicographical knowledge to ontologies in the Semantic Web.[51] The vocabulary has seen a number of updates and was published as a stable W3C vocabulary in May 2016.[52] This finalized version has since been adopted by a number of bodies, including the Global WordNet Association, to represent and link existing lexical resources on the Semantic Web.[53] It is this particular version that the name LEMON will henceforth be used to indicate.

LEMON consists of a number of modules. The core module, called OntoLex, contains terminology that is the most relevant for historical language thesauri. The main terms in this module are those for lexical entries (`LexicalEntry`), lexical senses (`LexicalSense`), forms (`Form`), and lexical concepts (`LexicalConcept`). Reminiscent of dictionaries – and in line with historical language thesauri – a lexical entry has one or more lexical senses and grammatical realizations, or forms. In order to organize lexical entries and senses not alphabetically but onomasiologically, i.e. by their meaning, it is possible to associate them with lexical concepts that can be organized hierarchically.

Lexical concepts in OntoLex are defined as a "mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses".[54] This definition, not unlike concepts in SKOS, appears highly suitable to express categories from historical language thesauri. In fact, lexical concepts from OntoLex are asserted to be specializations of SKOS concepts. The approach outlined in this chapter to use SKOS for the topical system of thesauri is therefore

[49] See Chapter 5, which introduces *lemon-tree*, a data vocabulary that contains terminology for expressing conceptual levels in thesauri.
[50] *OLiA Annotation Model for Uby Parts of Speech.*
[51] McCrae et al., 'Interchanging Lexical Resources on the Semantic Web'.
[52] 'Lexicon Model for Ontologies'.
[53] 'Global Wordnet Formats'.
[54] 'Lexicon Model for Ontologies'.

strengthened by the specification of LEMON, which makes this connection between the data vocabularies explicit.

As discussed, the LEMON vocabulary covers fundamental terminology for lexical senses in historical language thesauri and their relation to concepts. Next to coverage, the criteria for cohesion and standardization are also met by LEMON. Firstly, the asserted connection with SKOS enables cohesion and increases standardization and interoperability. Secondly, the LEMON data vocabulary has been popular since its creation and has continued to be widely supported. Notable resources that have so far been expressed in LEMON include Princeton's WordNet and Wiktionary,[55] Wikidata,[56] and FrameNet.[57] In short, the framework that LEMON offers is highly suitable for expressing lexical senses from historical language thesauri. Before delving into the matter on which data vocabularies best capture *attributes* of lexical senses, the next section will first discuss how the relation of synonymy, one of the main components found in thesauri, can be expressed.

3.6. Semantic Web form for the relation of synonymy

On the Semantic Web, a small number of available data vocabularies is capable of expressing synonymy as found in thesauri. Searches in *LOV* for 'synonym' and 'synonymy' located a number of candidate data vocabularies for this purpose. The majority of them contains terminology to capture synonymy between elements other than lexical senses.[58] However, as discussed in Chapter 1, the relation of synonymy as found in thesauri is one between senses.[59] Only three vocabularies define their synonymy relations as lexical relations or sense relations: LEMON, GOLD,[60] and LexInfo. The last two vocabularies are positioned as supplementing classifications to LEMON, useful for describing linguistic objects more thoroughly.[61] In order to satisfy the criterion of cohesion, a closer look at LEMON is therefore warranted, since this data vocabulary has already been selected for representing historical language thesauri (i.e., for expressing lexical senses, see section 3.5.1). If LEMON suffices in expressing synonymy for thesauri, its use is therefore preferred over that of GOLD and LexInfo.

In LEMON, the relation of synonymy is asserted not directly between senses but indirectly by grouping them under a lexical concept. Lexical senses that are associated with the same lexical concept through the `isLexicalizedSenseOf` relation are considered to be synonymous. The LEMON specification indicates how its terminology for synonymy corresponds to those used in lexical nets, such as University of Princeton's WordNet, which captures sets of synonyms (referred

[55] McCrae et al., 'Integrating WordNet and Wiktionary with *lemon*'.
[56] Nielsen, 'Lexemes in Wikidata'.
[57] Eckle-Kohler et al., 'lemonUby'.
[58] *LOV*, s.v. 'ru', 'scot', 'dbpedia-owl', 'uniprot'.
[59] Indeed, near-synonymy, which is a form of synonymy between lexical senses, has been called "the staple of thesauruses" (Murphy, 'Meaning Relations in Dictionaries', p. 448).
[60] In full, GOLD stands for General Ontology for Linguistic Description.
[61] 'Lexicon Model for Ontologies'.

to as synsets).[62] The specification thereby demonstrates that synonymy can be captured using solely this data vocabulary, which is shown suitable for capturing this important relation in historical language thesauri on the Semantic Web. With the relation of synonymy covered, this chapter will proceed to discuss which data vocabularies best capture sense attributes.

3.7. Semantic Web form for sense attributes

Most of the historical language thesauri of Scots and English contain information on lexical senses beyond merely the existence of these senses.[63] These thesauri specify such attributes as the part of speech, definition, language, and usage features. Although LEMON suggests definitions are best captured using the `definition` relation from SKOS,[64] choosing an appropriate data vocabulary for expressing the remainder of these attributes on the Semantic Web is not as straightforward as choosing that for the other components has been. Indeed, the LEMON specification states that it neither aims to provide such terminology itself nor wishes to proclaim a single data vocabulary as being the most suitable. Instead, the specification lists a number of efforts that may be useful for describing properties of linguistic objects: GOLD, LexInfo, OLiA, ISOcat (recently superseded by DatCatInfo), and the Clarin Concept Registry.[65] Thus, the specification fails to standardize sense attributes. The reluctance to put forward a single data vocabulary for this purpose is not without reason, as an examination of the expressivity of the existing data vocabularies will show.

3.7.1. Part of speech

The historical language thesaurus *TOE* classifies the verbs it contains as, simply, verbs or as intransitive or transitive verbs. In case of transitive verbs, this thesaurus leaves the number of objects each verb takes unspecified: they could be either monotransitive (and take only a single object) or ditransitive (and take both a direct and indirect object). When reviewing the terminology in the existing data vocabularies, it is not uncommon to find some of the aforementioned verb classifications missing. The GOLD vocabulary, for instance, contains terminology for intransitive verbs, monotransitive verbs, and ditransitive verbs, but lacks terminology for the more general notion of verbs and for transitive verbs of which it is left unspecified whether they are monotransitive or ditransitive. LexInfo, in contrast, has a generic term for verbs, but lacks discrimination in terminology between transitive and intransitive verbs. Some recent initiatives, such as that of the Global WordNet Association, have opted to create further data vocabularies to fulfil their specific needs in capturing parts of speech.[66] There, too, the terminology for parts of speech is incomplete, lacking

[62] Ibid.
[63] As demonstrated in Chapter 1.
[64] 'Lexicon Model for Ontologies'.
[65] Ibid.
[66] See 'Global Wordnet Formats'. The terminology that the association coined for parts of speech has been made available at http://globalwordnet.github.io/schemas/wn.

a distinction between transitive and intransitive verbs (not to mention an absence of such parts of speech as interjections, pronouns, and prepositions). DatCatInfo, employed by the Lexical Markup Framework, is a data category repository that contains definitions for parts of speech, amongst others, and constitutes a rich (albeit non-RDF) alternative to the data vocabularies mentioned above. The repository contains both coarse-grained and fine-grained distinctions for parts of speech, including on transitivity, but lists multiple definitions for a single part of speech, each with its own persistent identifier, and appears to be inconclusive as to which is preferred.[67]

It may well be that there is no one perfect data vocabulary for the parts of speech found across all historical language thesauri. Some of these vocabularies lack specific parts of speech, while others lack (or might even disagree with) the hierarchy between these parts of speech as employed by thesauri. Hans-Jürgen Diller's conclusion, mentioned in Chapter 1 in reference to the topical systems of thesauri, seems relevant for part of speech hierarchies, too: "[t]here is no one right classification; there are only more and less useful ones".[68] Which data vocabulary is best suited to express these particular sense attributes for a historical language thesaurus depends, therefore, on the exact needs of that thesaurus and is best approached on a case-to-case basis.

Regardless of the data vocabulary used, it is evident that the parts of speech together form a system to group lexical items. In that regard, parts of speech are similar to the topical system of thesauri, with two notable differences. Firstly, parts of speech are based on syntactic properties of such items instead of semantic ones. Secondly, they classify rather than categorize content because of the strict and clear criteria that the grouped items have to fulfil. The GOLD and LexInfo vocabularies acknowledge this fact by providing part of speech terminology in the form of a hierarchy of classes.

3.7.2. Usage features

Any attempt at finding a single data vocabulary to capture usage features of senses seems unrealistic. As Chapter 1 has shown, most usage features are indicated through labelling, but the meaning of each label (and how it relates to other labels) is specific to the thesaurus it is found in.[69] A label should therefore always be seen within the defined context of its body. As a result, a shared terminology and definitions for these labels is not likely to be found. LexInfo, for instance, defines a formal register but not an informal one. Moreover, the exact relation between its temporal qualifiers – `archaic`, `obsolete`, and `outdated` – is left unspecified and such relations may very well differ between thesauri. Are

[67] The repository contains four definitions for *verb*, of which two have the status "standardized". Further efforts beyond the Semantic Web towards standardizing parts of speech exist, too. A case in point is the Universal Dependencies framework, which employs the CoNLL-U file format to capture information on sentence tokens, including a universal part of speech tag (e.g., verb, noun) and a language-specific one (see 'CoNLL-U Format', *Universal Dependencies*).

[68] Diller, Review of *HTE1*, p. 322.

[69] Atkins and Rundell, *The Oxford Guide to Practical Lexicography*, pp. 182–6.

outdated items per definition also considered obsolete? Quite as is the case for the part of speech attribute, which vocabulary is best suited to express usage features may depend on the thesaurus in question. If no adequate vocabulary exists, some usage features may even best be represented with terminology specifically coined per thesaurus.

One usage feature that sometimes went beyond mere labelling in the existing historical language thesauri was diachronic marking: stating when a particular sense was in use. This temporal aspect is conveyed in thesauri through named, or even dated, periods in time. One particular data vocabulary, found in *LOV*, appears highly suitable for capturing such temporal aspects: the Time Ontology in OWL (OWL-Time).

OWL-Time for diachronic usage features

The OWL-Time vocabulary contains terminology for expressing temporal aspects and revolves around the notion of temporal entities (`TemporalEntity`).[70] A temporal entity can be either a point in time (`Instant`) or a period with a non-zero duration (`ProperInterval`). In addition to terminology to position such a temporal entity on a given calendar, OWL-Time also contains a number of terms to identify the relative ordering of temporal entities. These terms, based on work by James Allen, include relations to indicate that one interval has taken place before, during, or after another.[71] Thus, temporal aspects can be described quantitively (i.e., with their exact position on a timeline) or qualitatively (i.e., with their relative ordering).

The OWL-Time data vocabulary has been published by the W3C and is, as of June 2017, a candidate recommendation. The combination of LEMON and OWL-Time has already been explored on the Semantic Web. Anas Fahad Khan et al., for example, have employed both data vocabularies to express the temporal extent of lexical domains and semantic shifts.[72] Thus, the Old English period and subperiods thereof could be defined quantitatively (i.e., with exact dates) or qualitatively (i.e., in relation to other periods).[73] Isa Maks et al., too, have opted for this combination of data vocabularies for bringing diachronic lexicons to the Semantic Web.[74]

[70] 'Time Ontology in OWL'.

[71] Allen, 'Towards a General Theory of Action and Time'; and Allen and Ferguson, 'Actions and Events in Interval Temporal Logic'.

[72] Khan et al., 'Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web', pp. 42–3.

[73] Ibid., p. 44.

[74] Although the paper in the DHBenelux proceedings still shows a custom data vocabulary used for temporal usage features (Maks et al., "Integrating Diachronous Conceptual Lexicons through Linked Open Data"), the GitHub repository shows that the researchers transitioned to the OWL-Time vocabulary shortly after. https://github.com/cltl/clariah-vocab-conversion/tree/master/rdf-data.

3.7.3. Language

Capturing language, the final sense attribute covered in this chapter, has a more uniform approach than the part of speech and usage features do.[75] LEMON prescribes that the language attribute should be captured using a language tag. Such a tag consists of codes specifically meant to associate a string with a certain language and, possibly, with a specific country or region. The specification requires these language tags to be formed using the language codes based on ISO 639 and, optionally, a hyphen followed by an ISO 3166-1 country code. Thus, this practice requires the "en" tag for English or "en-GB" for English used in Great Britain specifically. Historical languages, too, can be expressed using these codes. Old English, for example, is identified by the language code "ang", Middle English by "enm", and Lowland Scots by "sc". The means to apply these tags to strings are inherent to the RDF format and do not depend on any specific data vocabulary.[76]

3.8. Conclusion

This chapter has discussed what form historical language thesauri should take on the Web. Already available forms of existing historical language thesauri either fall short in terms of the functionality they can offer or hamper reuse and interoperability of their content. A new digital form based on Semantic Web technology may improve the use and reuse of these lexicographic resources. The first step in establishing a Semantic Web form of historical language thesauri has been to select appropriate data vocabularies on the basis of three criteria: coverage, standardization, and coherency. Data vocabularies that meet these criteria have been located for many of the key components found in historical language thesauri of Scots and English. For the topical system, senses, and synonymy, SKOS and LEMON offer terminology that covers most needs. For other components of these thesauri, it has not been as straightforward to find already existing data vocabularies that meet the three criteria. For parts of speech, for instance, there may be no one right data vocabulary to use across all historical language thesauri. Be that as it may, the Semantic Web form constructed in this chapter is in line with prominent guidelines on publishing data on the Web, contributes to data being FAIR, and may well facilitate the functionality over historical language thesauri sought after by researchers — a hypothesis tested in Part III of this dissertation. Before this evaluation, the next two chapters of Part II will continue to discuss aspects of the Semantic Web form

[75] The language attribute is, of course, not solely applicable to lexical senses. Lexical entries, which group such senses, are language-specific, too.

[76] As with use of any registry, there are limitations of which one should be aware: codes registered might not be complete, sufficiently accurate, or sufficiently nuanced. Moreover, one may not be able to ascertain with certainty, especially in contexts of historical languages and the texts that have survived, whether a specific word belongs to a given language. Indeed, a word may even have been misread or misconstructed. Even so, language tags form a good starting point and appear suitable for use with the historical language thesauri listed in Chapter 1.

of historical language thesauri by identifying lacunae in LEMON and introducing complementary terminology for representing thesauri, specifically, in this form.

References

'AGROVOC Linked Data', *AIMS*. http://aims.fao.org/standards/agrovoc/linked-open-data. Accessed: 20 December 2017.

Allemang, D. and J. Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*, 2nd edn (Waltham, 2011).

Allen, J. F., 'Towards a General Theory of Action and Time', *Artificial Intelligence* 23 (1984), 123–54.

Allen, J. F. and G. Ferguson, 'Actions and Events in Interval Temporal Logic', in *Spatial and Temporal Reasoning*, ed. O. Stock (Dordrecht, 1997), pp. 205–45.

Atkins, B. T. S. and M. Rundell, *The Oxford Guide to Practical Lexicography* (Oxford, 2008).

Berners-Lee, T. et al., 'The Semantic Web', *Scientific American Magazine* (May 2001).

Berners-Lee, T., 'Linked Data: Design Issues', *W3C*. https://www.w3.org/DesignIssues/LinkedData.html. Last updated: 18 June 2009.

'Best Practices for Publishing Linked Data', eds. B. Hyland et al., *W3C*. http://www.w3.org/TR/ld-bp/. W3C Working Group Note. Created: 9 January 2014.

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

Cimiano, P. et al., *Linguistic Linked Data: Representation, Generation and Applications* (Cham, 2020). doi: 10.1007/978-3-030-30225-2.

'CoNLL-U Format', *Universal Dependencies*. https://universaldependencies.org/format.html.

'Data on the Web Best Practices', eds. B. F. Lóscio et al., *W3C*. http://www.w3.org/TR/DWBP/. W3C Recommendation. Created: 17 January 2017.

DatCatInfo. https://datcatinfo.net. Accessed: 20 July 2022.

Diller, H., Review of *HTE1*, *Anglia* 128.2 (2010), 319–23.

Eckle-Kohler, J. et al., 'lemonUby: A Large, Interlinked, Syntactically Rich Lexical Resource for Ontologies', *Semantic Web* 6.4 (2015), 371–8.

Encyclopedia of Language & Linguistics, 2nd edn, ed. K. Brown, 14 vols. (Oxford, 2006).

EuroVoc: Multilingual Thesaurus of the European Union. http://eurovoc.europa.eu/. Accessed: 20 December 2017.

General Ontology for Linguistic Description. http://purl.org/linguistics/gold/. Created: 2010.

'Getty Vocabularies', *The Getty Research Institute*. http://www.getty.edu/research/tools/vocabularies/. Accessed: 20 December 2017.

'Global Wordnet Formats', *Global WordNet Association*. http://globalwordnet.

github.io/schemas/. Accessed: 20 December 2017.

Hollink, L. et al., 'Thesaurus Enrichment for Query Expansion in Audiovisual Archives', *Multimedia Tools and Applications* 49.1 (2010): 235–57.

HTE1 = Historical Thesaurus of the Oxford English Dictionary, 2 vols., eds. C. Kay et al. (Oxford, 2009).

'HTML & CSS', *W3C*. http://www.w3.org/standards/webdesign/htmlcss. Accessed: August 15, 2017.

'Integrated Public Sector Vocabulary', *Joinup*. https://joinup.ec.europa.eu/catalogue/asset_release/integrated-public-sector-vocabulary. Accessed: 20 December 2017.

Isaac, A. and J. de Smedt, *ISO-THES*. http://purl.org/iso25964/skos-thes. Created: 17 March 2015.

'ISO 25964 Thesaurus Schemas', *NISO*. http://www.niso.org/schemas/iso25964/. Accessed: 13 June 2017.

Janowicz, K. et al., 'Why the Data Train Needs Semantic Rails', *AI Magazine* 36.1 (2015), 5–14.

Khan, A. F. et al., 'Tools and Instruments for Building and Querying Diachronic Computational Lexica', Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities, Osaka, December 2016, pp. 164–71. https://aclanthology.org/W16-4022.pdf.

Khan, A. F. et al., 'Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web', Proceedings of the 2nd International Workshop on Semantic Web for Scientific Heritage, Heraklion, 30 May 2016, pp. 37–45. https://ceur-ws.org/Vol-1595/paper4.pdf.

'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.

LexInfo, 2nd edn. http://www.lexinfo.net/ontology/2.0/lexinfo. Accessed: 20 December 2017.

Linked Open Vocabularies. https://lov.linkeddata.es/. Accessed: 13 June 2017 at http://lov.okfn.org/.

Maks, I. et al., "Integrating Diachronous Conceptual Lexicons through Linked Open Data", Proceedings of DHBenelux 2016, Belval, 8-10 June 2016, pp. 37–45.

McCracken, J., 'The Exploitation of Dictionary Data and Metadata', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 501–14.

McCrae, J. et al., 'Integrating WordNet and Wiktionary with *lemon*', in *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, eds. C. Chiarcos et al. (Heidelberg, 2012), pp. 25–34.

McCrae, J. et al., 'Interchanging Lexical Resources on the Semantic Web', *Language Resources and Evaluation* 46.4 (2012), 701–19.

'Medical Subject Headings', *U.S. National Library of Medicine*. https://www.nlm.nih.gov/mesh/. Accessed: 20 December 2017.

Murphy, M. L., 'Meaning Relations in Dictionaries: Hyponymy, Meronymy,

Synonymy, Antonymy, and Contrast', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 439–56.

MySQL 5.7 Reference Manual. https://dev.mysql.com/doc/refman/5.7/en/. Last updated: 19 December 2017.

'NASA Thesaurus', *NASA*. https://www.sti.nasa.gov/nasa-thesaurus/. Accessed: 20 December 2017.

Navigli, R., 'Word Sense Disambiguation: A Survey', *ACM Computing Surveys* 41.2 (2009), 1–69.

Nielsen, F., 'Lexemes in Wikidata: 2020 Status', Proceedings of the 7th Workshop on Linked Data in Linguistics, Marseille, May 2020, pp. 82-6. https://aclanthology.org/2020.ldl-1.12.pdf.

OLiA Annotation Model for Uby Parts of Speech. http://purl.org/olia/ubyCat.owl. Accessed: 20 December 2017.

'OWL 2 Web Ontology Language: Document Overview', 2nd edn, *W3C*. http://www.w3.org/TR/owl2-overview/. W3C Recommendation. Created: 11 December 2012.

'RDF/OWL Representation of WordNet', eds. M. van Assem et al., *W3C*. http://www.w3.org/TR/wordnet-rdf/. Last updated: 19 June 2006.

'RDF 1.1 Semantics', eds. P. J. Hayes and P. F. Patel-Schneider, *W3C*. https://www.w3.org/TR/rdf11-mt/. W3C Recommendation. Created: 25 February 2014.

'Semantic Web', *W3C*. https://www.w3.org/standards/semanticweb/. Accessed: 13 June 2017.

'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.

'Time Ontology in OWL', eds. S. Cox and C. Little, *W3C*. http://www.w3.org/TR/owl-time/. W3C Candidate Recommendation. Created: 6 June 2017.

'UNESCO Thesaurus', *UNESCO*. http://vocabularies.unesco.org/. Accessed: 20 December 2017.

Vandenbussche, P. et al., 'Linked Open Vocabularies: A Gateway to Reusable Semantic Vocabularies on the Web', *Semantic Web* 8.3 (2017), 437–52.

Van Assem, M., et al., 'A Method for Converting Thesauri to RDF/OWL', *International Semantic Web Conference* (Hiroshima, 2004), 17–31.

Van Assem, M. et al., 'A Method to Convert Thesauri to SKOS', *European Semantic Web Conference* (Budva, 2006), 95–109.

Van Assem, M., *Converting and Integrating Vocabularies for the Semantic Web*. Dissertation DPhil. VU University Amsterdam. 2010. https://research.vu.nl/ws/portalfiles/portal/42185333/complete+dissertation.pdf.

Wilkinson, M. D. et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data* 3.160018 (2016).

WordNet 3.0 in RDF, eds. M. van Assem and J. van Ossenbruggen. http://semanticweb.cs.vu.nl/lod/wn30/. Last updated: 25 September 2010.

'XKOS: An SKOS Extension for Representing Statistical Classifications', ed. F. Cotton, *DDI Alliance*. http://rdf-vocabulary.ddialliance.org/xkos.html. Unofficial Draft. Created: 18 January 2017.

129

3

Chapter 4

Information:
The paper that starts on the next page was published in 2017, available open access. The only change to that paper, here, is the inclusion of two numbers for the benefit of readers of the dissertation: the overall page number and the chapter number (presented in the margin in a grey box and a white box, respectively). When citing, please refer to the original publication and its page numbering.

Publication:
Sander Stolk, 'OntoLex and Onomasiological Ordering: Supporting Topical Thesauri', Proceedings of the LDK2017 Workshops, NUI Galway, 18 June 2017, pp. 60–67. http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_3.pdf.

OntoLex and Onomasiological Ordering: Supporting Topical Thesauri

Sander Stolk

Leiden University, Leiden, The Netherlands,
s.s.stolk@umail.leidenuniv.nl

Abstract. The OntoLex vocabulary has been designed to capture lexicons and to add their lexicographical knowledge to ontologies in the Semantic Web. Although the specification of the vocabulary posits that OntoLex allows lexicons to be ordered onomasiologically, it does so for a very specific kind of onomasiological ordering only. As a consequence, the vocabulary is currently insufficient for capturing a large proportion of the existing topical thesauri. This paper demonstrates the current expressivity and this shortcoming of OntoLex through two case studies: *The Historical Thesaurus of the Oxford English Dictionary* and *The Scots Thesaurus*. In order for OntoLex to offer full support for topical thesauri and their ordering principles, this paper proposes the addition of a single property to the vocabulary: `ontolex:isSenseIn`. ...

Keywords: OntoLex, Lemon, onomasiological ordering, thesaurus

1 Introduction

The Lexicon Model for Ontologies vocabulary has been designed to capture lexicons and to add their lexicographical knowledge to ontologies in the Semantic Web [1]. The vocabulary has seen a number of updates, and was published as a W3C vocabulary by the OntoLex community group in May 2016 [2]. This version, henceforth OntoLex, has since been picked up by a number of bodies, including the Global WordNet Association, to represent and link existing lexical resources on the Semantic Web [3].

The specification of OntoLex puts forward a manner in which "lexicons can be ordered onomasiologically, that is by meanings rather than by lemmas" [2]. For publishers of topical thesauri, this is good news indeed. Such support is essential for these lexicographical works, which order their words by meaning instead of from a to z as is common in typical dictionaries. Yet the OntoLex vocabulary supports a very specific kind of onomasiological ordering only. As a consequence, the vocabulary is currently insufficient for capturing the knowledge from a large proportion of the existing topical thesauri. The current paper demonstrates this shortcoming of OntoLex and proposes a way forward for the vocabulary.

2 Methodology

In order to provide insight into the current support of OntoLex for the onomasiological ordering of topical thesauri, this paper will present two case studies. The first is based on the *Historical Thesaurus of the Oxford English Dictionary* [4]; the second on *The Scots Thesaurus* [5]. Both lexicographical works employ an onomasiological ordering for their lexicon. The first-mentioned thesaurus is considered to be a distinctive one and contains sets of synonyms. The second is not distinctive but cumulative and refrains from indicating synonymy [6].

This paper expresses samples from both thesauri in the OntoLex vocabulary. The manner in which OntoLex is applied is in line with the specification of the vocabulary [2] and the approach outlined by the Global WordNet Association [3]. This approach has been adopted by several projects, amongst which the Open Dutch Wordnet [7]. Namespaces relevant for this paper are provided in Listing 1. The RDF snippets in subsequent listings are specified in the Turtle RDF syntax [8]. Sample data from the case studies correspond with resources between angular brackets in the RDF snippets (that is to say, their namespace is left unspecified for the present purpose).

Listing 1. Namespaces

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix wn: <http://wordnet-rdf.princeton.edu/ontology#> .
```

3 Case study *Historical Thesaurus of the OED*

The first case study presented here is that of the *Historical Thesaurus of the Oxford English Dictionary* (HTOED). HTOED captures the English lexis that has existed throughout its 1300-year history, from Old English up to Modern English. This topical thesaurus groups together lexical items that are considered near-synonymous and provides insight into their use in time and place. HTOED was first published in print in 2009 [4] and in the following year also electronically [9].

Figure 1 depicts a sample from HTOED. This sample contains six categories from the topical system of the thesaurus (here represented by circles), which are organized in a hierarchy. A category that is displayed lower than another category to which it is connected by means of a line is subordinate to that connected category. On the right, a number of lexical senses are displayed (some of which are obsolete, conveyed by a dagger sign). These senses are considered synonyms, or rather, near-synonyms, in HTOED and are members of the "Freedom/liberty" category.

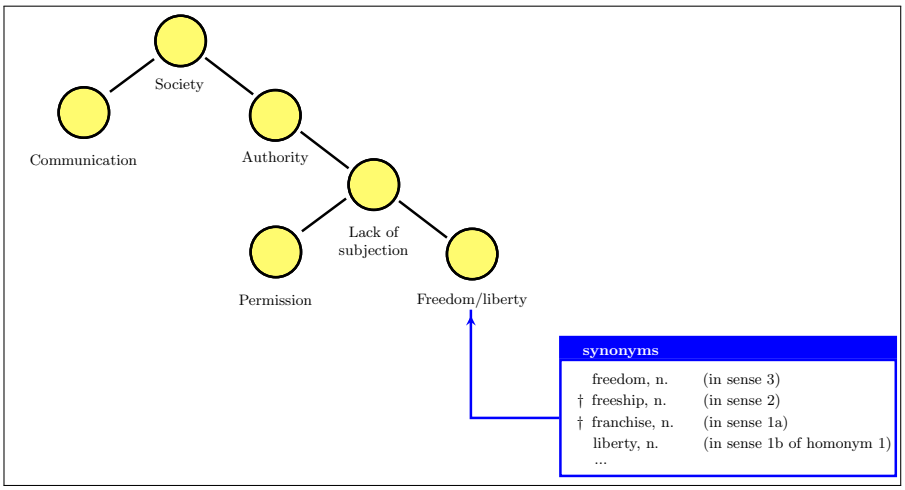


Fig. 1. Example HTOED content based on [9]

Expressing categories of the topical system of HTOED in OntoLex is relatively straightforward. Each HTOED category corresponds with a lexical concept in OntoLex. The latter is defined as a "mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses" [2]. This definition appears highly applicable to categories from topical thesauri. As lexical concepts are asserted to be specializations of SKOS concepts, it is possible to capture the hierarchy between categories using the `broader`/`narrower` relations from SKOS [10]. Listing 2 contains the RDF for expressing one of the HTOED categories in OntoLex, "Freedom/Liberty", and the relation to its superordinate category "Lack of subjection".

Listing 2. HTOED category "Freedom/liberty" expressed in OntoLex

```
<category-FreedomLiberty> a ontolex:LexicalConcept ;
    skos:prefLabel "Freedom/liberty"@en ;
    skos:broader <category-LackOfSubjection> .
```

The OntoLex vocabulary also contains terminology to express lexical senses and the lexical entries to which they belong. In order to state that a given lexical sense from HTOED belongs to one of its categories, the property `ontolex:isLexicalizedSenseOf` can be used. This property relates a lexical sense to a lexical concept, stating that it "lexicalizes" that concept. According to the section on Lexical Nets in the OntoLex specification, lexical senses that lexicalize the same concept are considered synonymous [2]. In other words, the relation of synonymy is not explicitly asserted in OntoLex, but can be inferred from the use of the `ontolex:isLexicalizedSenseOf` property. The resulting

RDF for the sense of *freedom* from the HTOED sample and its relation to the "Freedom/liberty" category is provided in Listing 3.

Listing 3. HTOED sense of *freedom* expressed in OntoLex

```
<sense-freedom-n-3> a ontolex:LexicalSense ;
    skos:prefLabel "freedom n. (sense 3)"@en ;
    ontolex:isSenseOf <entry-freedom-n> ;
    ontolex:isLexicalizedSenseOf <category-FreedomLiberty> .

<entry-freedom-n> a ontolex:LexicalEntry ;
    skos:prefLabel "freedom, n."@en ;
    wn:partOfSpeech wn:noun .
```

As shown, capturing the onomasiological ordering of the HTOED lexicon presents no issues with the OntoLex vocabulary. The vocabulary enables one to express categories and their hierarchy, lexical senses and their relation to a lexical entry, and the relation between the senses from HTOED and the categories to which they belong.

4 Case study *The Scots Thesaurus*

The second case study in this paper concerns *The Scots Thesaurus* (ScT) [5]. ScT captures the Lowland Scots lexis available throughout history, from its twelfth-century beginnings to the present. This thesaurus, published in 1990, categorizes its lexical items but does not indicate synonymy. Figure 2 depicts the sample taken from ScT, encompassing five categories and four lexical senses.

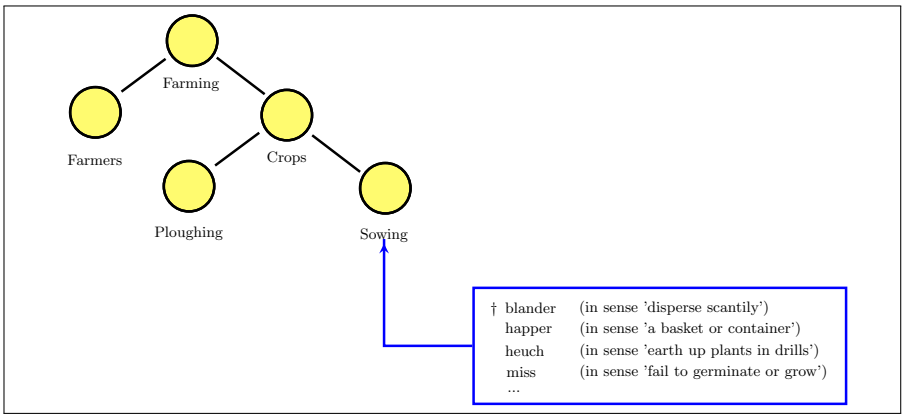


Fig. 2. Example ScT content

Expressing categories from ScT is possible in a manner identical to that used for HTOED. The result for the "Sowing" category from ScT, including its relation to the superordinate category "Crops", is provided in Listing 4.

Listing 4. ScT category "Sowing" expressed in OntoLex

```
<category-Sowing> a ontolex:LexicalConcept ;
    skos:prefLabel "Sowing"@en ;
    skos:broader <category-Crops> .
```

As for the lexical senses from ScT, these too can be expressed in OntoLex comparable to how it has been done for HTOED. There is, however, a notable difference. The property `ontolex:isLexicalizedSenseOf` is unsuitable for relating the senses of ScT to the categories to which they belong. The lexical senses in ScT are not necessarily lexicalizations of the category in question. Moreover, senses that belong to the same category are not necessarily considered synonymous. In fact, they rarely are. A case in point are the senses of *happer* and *miss* from the sample. Both of these senses are members of the category "Sowing", and indeed belong to that semantic domain, but can hardly be said to be synonymous or even to lexicalize the category.

What is missing, then, from the OntoLex vocabulary is terminology to express a looser manner of onomasiological ordering with categories than `ontolex:isLexicalizedSenseOf` does. The RDF snippet in Listing 5 contains the desired situation, where a tentative property `isSenseIn` is coined (see highlighted line) to express the relation between the sense of *blander* and the category to which it belongs.

Listing 5. ScT sense of *blander* expressed in OntoLex

```
<sense-blander-v-disperseScantily> a ontolex:LexicalSense ;
    skos:prefLabel "blander"@sco ;
    skos:definition "disperse scantily"@en ;
    ontolex:isSenseOf <entry-blander-v> ;
    :isSenseIn <category-Sowing> .

<entry-blander-v> a ontolex:LexicalEntry ;
    skos:prefLabel "blander, v."@sco ;
    wn:partOfSpeech wn:noun .
```

In short, OntoLex itself does not yet provide terminology to onomasiologically order the lexicographical content of ScT – and of other thesauri like it.

5 Discussion

The two case studies have shown that OntoLex is not yet expressive enough to indicate the relation between senses and categories for all topical thesauri. In fact,

the lack of a property like the tentative `isSenseIn` does not just affect conveying content from ScT and the great many existing cumulative thesauri like it. It also affects expressing these very relations found in thesauri such as HTOED. After all, senses in HTOED are not just lexicalizations of a category, they are also members of a number of categories. To illustrate, the assertion that the HTOED sense of *freedom* is a lexicalization of the category "Freedom/liberty" entails that this sense is a member of not just that category but also of its superordinate categories (see Listing 6).

Listing 6. HTOED sense of *freedom* and its relation to the categories of HTOED

```
<sense-freedom-n-3> a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf <category-FreedomLiberty> ;
    :isSenseIn <category-FreedomLiberty> ,
               <category-LackOfSubjection> ,
               <category-Authority> ,
               <category-Society> .
```

In order to truly express how senses are categorized according to topical systems in thesauri, then, additional terminology is required beyond what OntoLex currently offers. Properties from other vocabularies that might fill the gap, such as the `subject` property from Dublin Core Terms [11], tend to be too generic to be able to infer further knowledge from topical systems of thesauri. Moreover, the relation between such properties and `ontolex:isLexicalizedSenseOf` is not evident. As such, the required terminology is best captured in an update of the OntoLex vocabulary itself. The small addition of a single property such as `isSenseIn` (see Listing 7), then, and asserting its connection to the existing OntoLex property (see Listing 8) would enable onomasiological ordering of lexicons in topical thesauri of all varieties – distinctive or cumulative, and regardless of whether synonymy is indicated between senses.

Listing 7. Suggested OntoLex property `isSenseIn`

```
ontolex:isSenseIn a owl:ObjectProperty ;
    rdfs:label "is sense in"@en ;
    rdfs:comment "This property relates a lexical sense to a
                  concept that captures its meaning to some
                  extent (that is, partially or even fully)."@en ;
    rdfs:domain ontolex:LexicalSense ;
    rdfs:range ontolex:LexicalConcept .
```

Listing 8. Connection between existing OntoLex property and the suggested one

```
ontolex:isLexicalizedSenseOf
   rdfs:subPropertyOf ontolex:isSenseIn .
```


6 Conclusion

This paper has shown, by means of two case studies, to what extent the OntoLex vocabulary currently supports relating lexical senses to the concepts that facilitate an onomasiological ordering. Such an ordering is (by their very definition) used in lexicographical works known as topical thesauri. As it stands, the OntoLex vocabulary offers some support for those thesauri considered to be distinctive and that capture synonymy. Such thesauri ensure that lexical senses displayed at a certain category do not just belong to that category, but also express (or lexicalize) that category. Those thesauri that do not have that same level of specificity, but merely use their categories to organize lexical senses into semantic domains, are not yet supported by the terminology in OntoLex.

The small addition of a single property, as suggested in this paper, would have a big impact on the expressivity of OntoLex. The onomasiological ordering of both distinctive and cumulative thesauri – regardless of whether these thesauri indicate synonymy – could then properly be conveyed on the Semantic Web. As a result, the variety of lexicographical resources that sit comfortably in OntoLex would not be limited to dictionaries and lexical nets, as is presently the case, but would also include thesauri. Increased support in OntoLex for onomasiological ordering, then, would allow all these resources to truly shine on the Web. In short, ordering by meaning through the new `ontolex:isSenseIn` is both meaningful and sensible.

References

1. McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gmez-Prez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation **46**(4), 701–719 (2012)
2. Lexicon Model for Ontologies: Community report, 10 May 2016 (2016). URL http://www.w3.org/2016/05/ontolex/
3. Global WordNet Association: Global Wordnet formats. URL http://globalwordnet.github.io/schemas/
4. Kay, C., Roberts, J., Samuels, M., Wotherspoon, I. (eds.): Historical thesaurus of the Oxford English Dictionary: with additional material from "A thesaurus of Old English". Oxford University Press, Oxford (2009)
5. Macleod, I., Cairns, P., Macafee, C., Martin, R. (eds.): The Scots thesaurus. Aberdeen University Press, Aberdeen (1990)

6. Kay, C., Alexander, M.: Diachronic and synchronic thesauruses. In: P. Durkin (ed.) The Oxford handbook of lexicography, pp. 367–380. Oxford University Press, Oxford (2016)
7. Postma, M., van Miltenburg, E., Segers, R., Schoen, A., Vossen, P.: Open Dutch WordNet. In: Proceedings of the Eighth Global Wordnet Conference. Bucharest, Romania (2016)
8. Beckett, D., Berners-Lee, T., Prud'hommeaux, E., Carothers, G.: RDF 1.1 Turtle: W3C recommendation 25 February 2014 (2014). URL http://www.w3.org/TR/turtle/
9. Historical thesaurus of the Oxford English Dictionary (2010). URL http://oed.com/thesaurus
10. SKOS Simple Knowledge Organization System reference: W3C recommendation 18 August 2009 (2009). URL http://www.w3.org/TR/skos-reference/
11. DCMI metadata terms (2012). URL http://purl.org/dc/terms/

Chapter 5

Information:
The paper that starts on the next page was published in 2019, available open access under the CC BY license. The only change to that paper, here, is the inclusion of two numbers for the benefit of readers of the dissertation: the overall page number and the chapter number (presented in the margin in a grey box and a white box, respectively). When citing, please refer to the original publication and its page numbering.

Publication:
Sander Stolk, '*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.

lemon-tree: Representing Topical Thesauri on the Semantic Web

Sander Stolk
Leiden University, Leiden, The Netherlands
s.s.stolk@hum.leidenuniv.nl

Abstract

An increasing number of dictionaries are represented on the Web in the form of linguistic linked data using the *lemon* vocabulary. Such a representation facilitates interoperability across linguistic resources, has the potential to increase their visibility, and promotes their reuse. Lexicographic resources other than dictionaries have thus far not been the main focus of efforts surrounding *lemon* and its modules. In this paper, fundamental needs are analysed for representing topical thesauri specifically and a solution is provided for two important areas hitherto problematic: (1) levels that can be distinguished in their topical system and (2) a looser form of categorization than lexicalization. The novel *lemon-tree* model contains terminology to overcome these issues and acts as bridge between existing Web standards in order to bring topical thesauri, too, to the Semantic Web.

2012 ACM Subject Classification Information systems → Semantic web description languages; Information systems → Thesauri

Keywords and phrases lemon-tree, lemon, OntoLex, SKOS, thesaurus, topical thesaurus, onomasiological ordering, linked data

Digital Object Identifier 10.4230/OASIcs.LDK.2019.16

1 Introduction

An increasing number of dictionaries are represented on the Web in the form of linguistic linked data using the *lemon* vocabulary (e.g. [3, 12]). Such a representation facilitates interoperability across linguistic resources, has the potential to increase their visibility, and promotes their reuse [5, 13]. The core of the *lemon* vocabulary, OntoLex, has been designed to capture lexicons and to add their lexicographical knowledge to ontologies on the Web [14]. As capturing lexicographic information was not part of the primary aim of OntoLex, recent modules for *lemon* have sought to improve support for expressing such information [12, 2]. Using these modules, content of lexicographic resources can become part of the Linguistic Linked Data Cloud whilst minimizing information loss in the transition [2]. These modules, however, have explored mainly the need to represent dictionaries but not other lexicographical works such as topical thesauri. Indeed, previous research points out that additional terminology is needed for such thesauri [21]. The current paper aims to fill this gap by putting forward a novel model for this purpose: *lemon-tree*.

A topical thesaurus is a lexicographical work that organizes its lexical items according to their meaning (rather than alphabetically) by means of a topical structure [7, 9]. This overarching structure offers generic meanings to users as a starting point, which branch out to meanings increasingly specific. Once users locate the meaning which they are interested in, they are presented with the words or phrases that express that meaning. This overarching topical system in a thesaurus thus allows the user to move from meaning to lexical item [8].

The new *lemon-tree* vocabulary, described in this paper, bridges the existing standards SKOS [19] and *lemon* in order to express the content of topical thesauri on the Web. The SKOS vocabulary already allows for sharing concepts in RDF and organizing them in hierarchies. The *lemon* model and its core module OntoLex allow for sharing lexical entries, senses, and further lexicographic material. Terminology from both the SKOS and *lemon*

© Sander Stolk;
licensed under Creative Commons License CC-BY
2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 16; pp. 16:1–16:13
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

standards, then, are valuable for sharing topical thesauri on the Web in an interoperable manner. The *lemon-tree* model therefore aims to facilitate their combined use for that purpose, adding some terminology for perceived lacunae.

2 Methodology

In order to provide insight into fundamental needs for representing topical thesauri on the Web beyond those for other lexicographic material (e.g., dictionaries), this paper will explore elements specific to the structure of topical thesauri. For each such element or structuring, the extent is discussed with which SKOS and *lemon* OntoLex offer terminology to represent these elements. For lacunae, available terminology in the new *lemon-tree* model is discussed that is fit for the purpose. Each topic is illustrated by means of an existing thesaurus that exemplifies the matter at hand. Listed in order of their appearance, these thesauri are:

- Historical Thesaurus of the Oxford English Dictionary (HTOED) [10]
- Shakespeare Thesaurus (ShT) [20]
- Scots Thesaurus (ScT) [15]
- Love, Sex, and Marriage (LSM) [4]
- Roget's Thesaurus (Roget's) [18]

Figure 1 is a legend to the images in this paper that depict the content of existing thesauri.

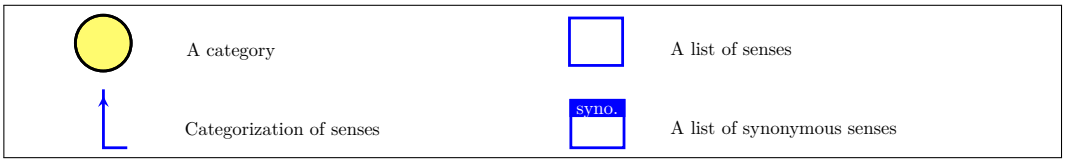


Figure 1 Legend.

Namespaces of the vocabularies relevant for this paper are provided in Listing 1. The RDF snippets in subsequent listings are specified in the Turtle RDF syntax [1]. In these snippets, samples taken from existing thesauri correspond with resources between angular brackets (that is to say, their namespace is left unspecified for the present purpose).

Listing 1 Namespaces.

```
@prefix tree: <https://w3id.org/lemon-tree#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix skos: <http://www.w3.org/2004/02/skos#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

Before going to the analysis proper, the next section will first provide a short background on topical thesauri. The section that follows treats the topical system, along with the different kinds of levels distinguished in such a system. Afterwards, words and their place within the topical system are discussed, followed by the conclusion.

3 Topical thesaurus

A topical thesaurus is a lexicographic resource that organizes its items according to their meaning rather than alphabetically [7, 9]. They do this by means of a topical structure: a tree of concepts. This overarching structure offers generic meanings to users as a starting point, which branch out to meanings increasingly specific. Once users locate the meaning which they are interested in, they are presented with the words or phrases that express that meaning. This overarching topical system in a thesaurus thus allows the user to move from meaning to lexical item. Figure 2 displays the main components of such a thesaurus, using a sample of the Historical Thesaurus of the Oxford English Dictionary [10]. The senses of four nouns are shown to be categorized under "Freedom/liberty" (of which those marked with a cross no longer exist). As these four senses convey the same meaning, they are thought to be loosely synonymous.

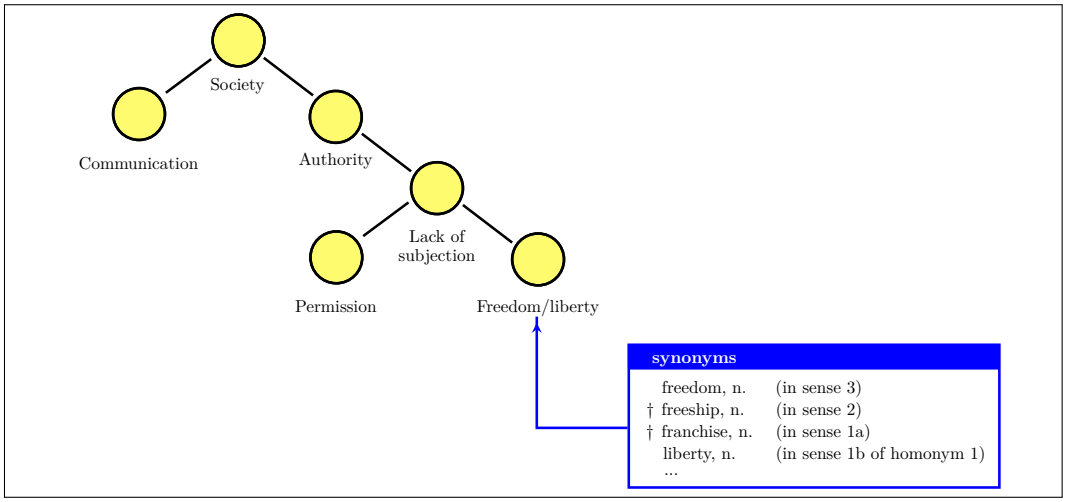


Figure 2 Thesaurus components, based on [11].

In a topical thesaurus, then, a word or phrase in a specific sense is located (or categorized) within a topical system, may be part of a set of synonyms, and is typically accompanied by additional lexicographic information such as its part of speech and usage features.

4 Topical system

The topical system of a thesaurus is its overarching structure used to organize lexical items. This structure is not unlike the taxonomies of animals and plants created by the eighteenth-century biologist Carl Linnaeus (1707-1778) and later expanded by Georges Cuvier (1769-1832) [6]. In these tree-like structures, the most generic or abstract concepts are used as roots, which branch out to concepts increasingly specific in meaning. Such topical systems can be represented with terminology from SKOS. Indeed, this standard from W3C was designed specifically for knowledge organization systems, including topical systems. Thus, the topical system as a whole would be captured as follows for the Historical Thesaurus of the Oxford English Dictionary.

Listing 2 A topical system in lemon-tree.

```
<htoed> a skos:ConceptScheme ;
    skos:prefLabel "Historical Thesaurus of the
                    Oxford English Dictionary"@en .
```

Its category "Freedom/liberty" can be captured as a SKOS **Concept**, part of the **ConceptScheme** of the topical system, and with its relation to its parent category "Lack of subjection" made explicit.

Listing 3 A category in lemon-tree.

```
<freedom-liberty> a skos:Concept ;
    skos:prefLabel "Freedom/liberty"@en ;
    skos:inScheme <htoed> ;
    skos:broader <lack-of-subjection> .
```

As we will see further on in the document, it is possible to use a specialized variant of SKOS Concept when categorizing senses. This topic will be treated in the section "Categorization and lexicalization".

4.1 Levels and depth

In a topical system, much like in any tree data structure, it is possible to distinguish multiple levels. Each level is found at a specific depth. For thesauri, however, there tend to be two forms of levels. Their topical system, after all, is meant to capture meaning and can therefore be subdivided into both levels of the tree structure and levels of meaning: tree levels and conceptual levels. SKOS and *lemon* OntoLex do not yet provide adequate terminology to capture these two levels and to distinguish them from another. The following subsections will discuss each of these levels in more detail and provides examples of how *lemon-tree* can be used to represent them on the Web.

4.1.1 Tree levels

A topical system of a thesaurus consists of categories that have been placed in a hierarchy. This hierarchical structure can be described using words for data structures known as trees. Each category in the hierarchy is a *node* in the tree, the nodes at the very top of the tree are called *roots*, and relations between nodes are known as *edges*. Each node is positioned at a certain *depth* of the tree. *Roots*, part of the first tree level, are at depth 0; nodes positioned directly below a root are at depth 1; nodes directly below these are at depth 2, and so on. Figure 3 displays such tree levels for the topical system of Roget's Thesaurus [18], perhaps the most well-known topical thesaurus in existence. Categories displayed on the same dotted line are part of the same tree level. Thus, the categories "Abstract Relations" and "Voluntary Powers" are part of the first tree level, at depth 0.

Tree levels can, of course, be calculated from the position of each node in the tree structure. Even so, some communities find it worthwhile to capture this information explicitly, too. Indeed, terminology to represent tree levels can already be found in XKOS, a vocabulary that extends SKOS [22]. In XKOS, each tree level is seen as a collection of categories, positioned at a specific tree depth. These collections are specialized SKOS **Collection**s. Although XKOS can capture tree levels in the topical system of a thesaurus, it cannot be used to capture its conceptual levels.

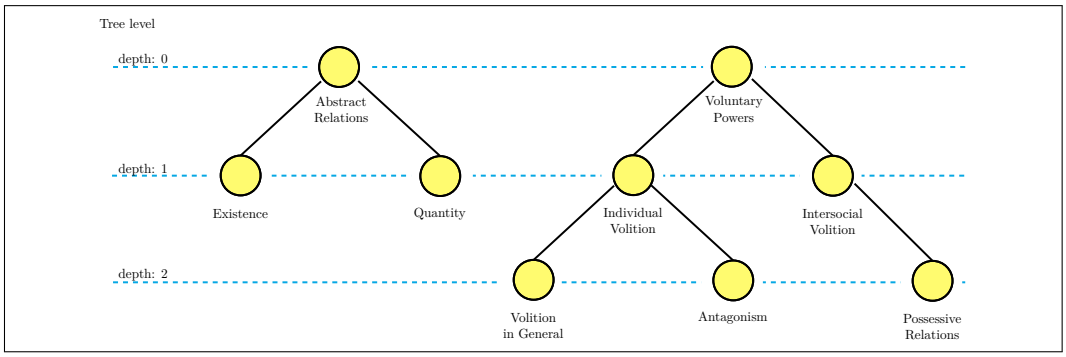


Figure 3 Tree levels (Roget's).

4.1.2 Conceptual levels

Next to tree levels, Roget's Thesaurus also contains conceptual levels. The thesaurus provides an outline of its topical system, which includes clear distinctions posited by its editor: categories in Roget's are not simply known as categories, but go by the name of *class*, *division*, or *section*. Indeed, the topical system starts out with six of these *classes*, which may branch out into *divisions* which are more specific, and ultimately into *sections*. A sample of its contents that includes these names is shown in Figure 4.

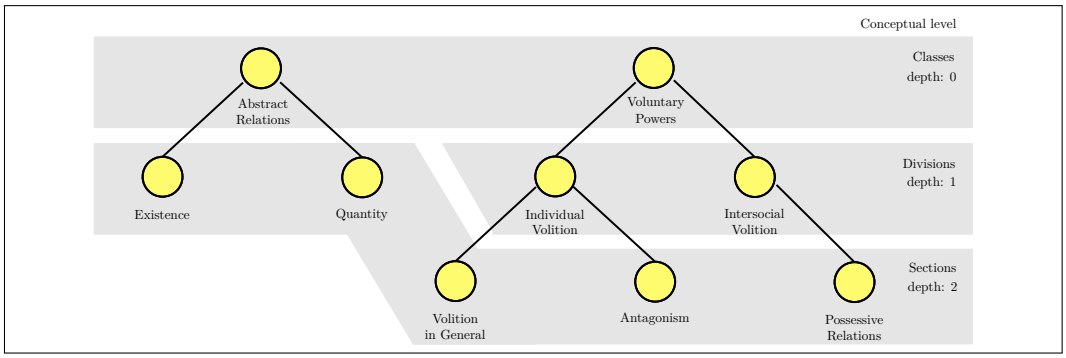


Figure 4 Conceptual levels (Roget's).

It is plain to see that the three types of category in Roget's act as a level of sorts. *Classes*, such as "Abstract Relations" and "Voluntary Powers", convey the highest level of abstraction; *sections* convey the lowest. Intuitively, categories of a higher level of abstraction branch out only to categories of a lower level of abstraction. As a consequence, we do not find categories known as *sections* in Roget's Thesaurus branching out into *classes* or *divisions*.

These levels mentioned do not necessarily map one-to-one with tree levels. In Figure 4, for example, both *divisions* and *sections* may be part of the 2nd tree level (at tree depth 1). Other thesauri, too, use similar notions to distinguish such conceptual levels [10] [4]. In the Historical Thesaurus of the Oxford English Dictionary, the first conceptual level consists of *sections*, followed by *categories* and lastly *subcategories*. Here, unlike in Roget's Thesaurus, a single category can branch out to categories from both the same conceptual level and one level beyond. A case in point is "Freedom/liberty". This is one of the so-called *categories* and branches out to a number of other *categories* (including "Independence" and "Liberation") but also to *subcategories* (including "Civil liberty" and "Moral freedom").

The *lemon-tree* model offers terminology to express these conceptuals levels. Although these levels are different from tree levels, the patterns in which thee former are captured in *lemon-tree* are analogous to how tree levels are captured in XKOS: a **ConceptualLevel** represents the level, the **conceptualDepth** property is used to indicate the depth of a level and **conceptualLevels** provides a means to list all available levels. The definitions below will be followed by snippets in which these three terms are employed.

▶ **Definition 1.** *ConceptualLevel (Class)*
A collection of concepts which are considered to be at the same conceptual depth (that is, semantically distanced from the root node). This conceptual depth may for certain thesauri coincide with the tree depth, but that is not necessarily the case for all thesauri.

SubClassOf: *skos:Collection*

▶ **Definition 2.** *conceptualDepth (DatatypeProperty)*
The depth of the conceptual level that groups a number of concepts. The conceptual depth in thesaurus taxonomies can only increase in a branch, but never decrease. The first conceptual level in a thesaurus is at depth 0; the next one at depth 1, etc.

Domain: *ConceptualLevel* ∪ *skos:Concept*
Range: *xsd:integer*

▶ **Definition 3.** *conceptualLevels (ObjectProperty)*
Provides the list of conceptual levels for a concept scheme.

Domain: *skos:ConceptScheme*
Range: *rdf:List*

Listing 4 A conceptual level in lemon-tree (Roget's).

```
<sections> a tree:ConceptualLevel ;
    skos:prefLabel "Sections"@en ;
    tree:conceptualDepth 2 ;
    skos:member <existence> ;
    skos:member <quantity> ;
    skos:member <volition-in-general> ;
    skos:member <antagonism> ;
    skos:member <possessive-relations> .

<rogets> a skos:ConceptScheme ;
    skos:prefLabel "Roget's Thesaurus"@en ;
    tree:conceptualLevels ( <classes> <divisions> <sections> ) .
```

Listing 5 A conceptual level in lemon-tree (HTOED).

```
<categories> a tree:ConceptualLevel ;
    skos:prefLabel "Categories"@en ;
    tree:conceptualDepth 1 ;
    skos:member <freedom-liberty> ;
    skos:member <lack-of-subjection> ;
    skos:member <permission> ;
    skos:member <authority> ;
    skos:member <communication> ;
    skos:member <society> ;
    skos:member <independence> ;
    skos:member <liberation> .

<htoed> a skos:ConceptScheme ;
    skos:prefLabel "Historical Thesaurus of the
                    Oxford English Dictionary"@en ;
    tree:conceptualLevels
                    ( <sections> <categories> <subcategories> ) .
```


The next section will discuss words and their place within the topical system.

5 Words and senses

A thesaurus contains lexical items that have been categorized, allowing users to go from meaning to words or phrases that express that meaning. Such words and senses can be represented using *lemon* OntoLex terminology. A word or phrase is captured as an OntoLex LexicalEntry and each of its senses as a LexicalSense. Examples thereof are presented below.

Listing 6 A lexical entry in lemon-tree.

```
<entry-freedom> a ontolex:LexicalEntry ;
    rdfs:label "freedom"@en ;
    ontolex:canonicalForm [
        a ontolex:Form ;
        ontolex:writtenRep "freedom"@en ;
    ] .
```

Listing 7 A lexical sense in lemon-tree.

```
<sense-freedom-3> a ontolex:LexicalSense ;
    ontolex:isSenseOf <entry-freedom> .
```

For further details on the notion of LexicalEntry and LexicalSense, we refer the reader to the *lemon* documentation. Advice on how to best capture other aspects of lexical items (e.g., their part of speech and other labels) is provided there, too.

5.1 Categorization

Topical thesauri do not categorize lexical items or word-forms but lexical senses: words or phrases in a particular sense. This statement may at first glance appear counter-intuitive for users of thesauri, as a number of these resources simply present head-forms of a word (or phrase) as member of their categories. In the Shakespeare Thesaurus [20], for instance, category "01.02 sky" contains the following item:

> **heaven, n.**

The head-form "heaven" in this example is similar in appearance to a headword, or lemma, found in typical dictionaries. This gives off the appearance that thesauri categorize lexical items. The following fictitious dictionary entry, however, demonstrates otherwise.

> **heaven, n.** 1) abode of one or more gods 2) the sky

It is evident that the "heaven, n." entry in the Shakespeare Thesaurus, found in the category "01.02 sky", represents the lexical item heaven in not all of its senses listed above but in only the second sense.

Werner Hüllen, who has thoroughly researched the topical tradition of thesauri, confirms that the entries in thesauri represent senses [8]. Further confirmation that topical thesauri categorize lexical senses can be found in the online edition of the Historical Thesaurus of the Oxford English Dictionary [11]. This edition takes advantage of both the topical structure of the thesaurus and the full dictionary entries of the Oxford English Dictionary. This rich set-up allows for a closer investigation of the relation between a thesaurus and entries in a dictionary. Dictionary entries in the Oxford English Dictionary have a number of senses. Each sense listed contains a reference to a thesaurus category. Conversely, the thesaurus categories in this edition list the senses they contain and provide hyperlinks not simply to dictionary entries but to specific senses *within* these entries. As such, it is evident that this thesaurus indeed categorizes senses of lexical entries, and not lexical entries as a whole. In the next section, we will provide more detail on categorization and how to capture it using *lemon-tree*.

5.2 Categorization and lexicalization

There are two forms of categorization to be found in thesauri. In the Historical Thesaurus of the Oxford English Dictionary, words in a particular sense directly express their concept. These words are said to *lexicalize* that concept. In Figure 2, "freedom" and "liberty" can directly be used if one wants to express "Liberty/freedom". This used to be the case for "freeship" and "franchise", too, in the history of the English language. (As this is no longer the case, these word senses are marked with a cross in front of them.)

Such lexicalization is not present in every thesaurus, however. In fact, it is more often the case than not in thesauri that it is absent. The sample in Figure 5 has been taken from the Scots Thesaurus [15] and illustrates this lack of lexicalization. Here, the sense 'to disperse scantily' of "blander" can hardly be said to directly express "Sowing". This is likewise the case for the sense 'a basket or container' of "happer". These senses may have a relation to the concept of "Sowing" but they do not lexicalize that concept. Their meaning causes them to be listed as part of the concept instead, that they are senses *in* that concept, as it were. Note that senses that lexicalize a concept are by definition senses also found *in* that concept. In other words, lexicalization is a special form of categorization.

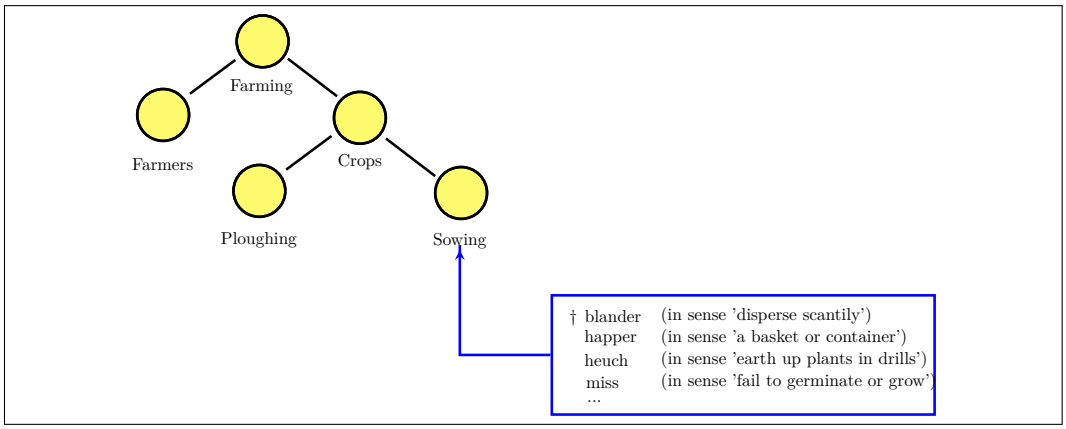


Figure 5 Sample from The Scots Thesaurus.

For asserting that senses are lexicalizations of a concept, OntoLex offers the property **isLexicalizedSenseOf**. For categorization, however, current vocabularies do not offer terminology expressive enough to capture the distinction with lexicalization and the connection between these two relations [21]. A case in point is the OntoLex property **reference**, which might appear suitable at first glance. Indeed, the property allows referring to a concept from a lexical sense. There are two problems with its use in the context of topical thesauri, however. Firstly, there is no mention in OntoLex of any direct relation between **isLexicalizedSenseOf** and **reference**, which leaves the important connection between lexicalization and categorization unexpressed and hinders inferring further knowledge from topical systems of thesauri. Secondly, the property **reference** is considered a functional one. As such, a sense may reference a single concept only using this property. However, when a sense in a thesaurus is categorized as part of a given concept, that sense is automatically also categorized as part of any parent concepts (i.e., the sense of "blander" in The Scots Thesaurus is categorized not just with "Sowing" but also with "Crops", and "Farming"). A functional characteristic therefore does not fit in this context. Instead, *lemon-tree* offers the property **isSenseInConcept** to capture these nuances needed for topical thesauri.

▶ **Definition 4.** *isSenseInConcept (ObjectProperty)*
This property relates a lexical sense to a concept that captures its meaning to some extent (that is, partially or even fully).

SubPropertyOf: *dcterms:subject*
Domain: *ontolex:LexicalSense*
Range: *skos:Concept*

The relation between **isSenseInConcept** and terminology from Ontolex has been added to the Lemon-tree model. As a result, the Ontolex property **isLexicalizedSenseOf** is asserted to be a sub property of **isSenseInConcept**. Moreover, the property **evokes** has an additional property chain of Ontolex **sense** followed by **isSenseInConcept**.

▶ **Definition 5.** *ontolex:isLexicalizedSenseOf (ObjectProperty)*

SubPropertyOf: *isSenseInConcept*
Domain: *ontolex:LexicalSense*
Range: *ontolex:LexicalConcept*

▶ **Definition 6.** *ontolex:evokes (ObjectProperty)*

Domain: *ontolex:LexicalEntry*
Range: *ontolex:LexicalConcept*
PropertyChain: *ontolex:sense o isSenseInConcept*

The examples below show how both categorization and lexicalization can be captured by employing the properties **isSenseInConcept** and **isLexicalizedSenseOf**. Notice that the property to express lexicalization is used in the example of the Historical Thesaurus of the Oxford English Dictionary. There, the use of this property automatically indicates that the category is not only a SKOS **Concept**, but a concept that is expressed or lexicalized. Such a concept is called a **LexicalConcept** according to OntoLex.

Listing 8 Categorization in lemon-tree.

```
<sense-happer-basket> a ontolex:LexicalSense ;
    tree:isSenseInConcept <sowing> .

<sowing> a skos:Concept .
```

Listing 9 Lexicalization in lemon-tree.

```
<sense-freedom-3> a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf <freedom-liberty> .

<freedom-liberty> a ontolex:LexicalConcept .
```

It should be noted that, whenever definitions are available for senses, it is possible to make these definitions part of the topical system. After all, the topical system allows a user to go from meaning to items that express that meaning. A sense definition is just such a meaningful item. The snippet below shows the result of this practice when applied to The Scots Thesaurus. Here, an additional concept is added to the topical system. This concept represents the sense definition of “happer” and is lexicalized by this sense.

Listing 10 Sense definitions as concepts.

```
<sense-happer-basket> a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf <a-basket-or-container> .

<a-basket-or-container> a ontolex:LexicalConcept ;
    skos:prefLabel "a basket or container"@en
    skos:broader <sowing> .

<sowing> a skos:Concept .
```

This approach has a caveat: synonyms are expected to lexicalize the same concept. Existing thesauri may not contain information for this additional level of grouping, requiring additional efforts in their transition to a Semantic Web form.

5.3 Synonymy

Categories in a topical system group lexical senses into sets with a similar or related meaning. In some thesauri, though certainly not all, sets exist that indicate an even stronger semantic tie: one of synonymy. A case in point is the Historical Thesaurus of the Oxford English Dictionary, in which senses placed at the same category are deemed loosely synonymous. That is to say, grouped senses in this thesaurus have a similarity in meaning and are interchangeable in specific contexts. The introduction to the thesaurus Love, Sex, and Marriage discusses synonymy found in thesauri as follows: [4]

> Grouping terms together in a thesaurus, even in a thesaurus as detailed as this, does not imply absolute synonymy. Many scholars doubt whether absolute interchangeability is actually possible.

Instead of absolute synonymy, then, it is common to find a looser form of synonymy in thesauri. This form is referred to as near-synonymy [16].

Near-synonymy is evident for lexical senses that lexicalize the same concept. After all, such senses directly express the same meaning. Thus, all the senses that lexicalize category "Freedom/liberty" of the Historical Thesaurus of the Oxford English Dictionary are known to be near-synonyms. Thus, synonymy can already be captured using terminology from *lemon* OntoLex. Using further vocabularies, it is also possible to link synonyms together via a direct relation between LexicalSenses, or to form groups of synonyms known as synsets if so desired [14].

Listing 11 Synonymy in lemon-tree.

```
<sense-freedom-3> a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf <freedom-liberty> .
<sense-freeship-2> a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf <freedom-liberty> .
<sense-franchise-1a> a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf <freedom-liberty> .
<sense-liberty1-1b> a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf <freedom-liberty> .
```

6 Conclusion

This paper set out to analyse fundamental needs for representing topical thesauri on the Web and to supply a solution for problematic areas encountered. The standardized SKOS and *lemon* vocabularies have shown to be of great value in expressing the topical system and lexical items in such a thesaurus respectively. There are, however, a few important aspects in which they fall short. The most notable two are: (1) levels that can be distinguished in a topical system and (2) a looser form of categorization than lexicalization. The novel *lemon-tree* model contains terminology to fill this gap and acts as bridge between the existing Web standards. As this paper has demonstrated, *lemon-tree* allows capturing a variety of topical thesauri –

each with its own particular characteristics. Indeed, the model has thus far been employed successfully in transitioning A Thesaurus of Old English [17] to linguistic linked data and has been found to be a good fit for the Mittelhochdeutsche Begriffsdatenbank [23]. The full data model of *lemon-tree* and its specification can be found at `https://w3id.org/lemon-tree#`.

References

1 David Beckett, Tim Berners-Lee, Eric Prud'hommeaux, and Gavin Carothers. RDF 1.1 Turtle: W3C recommendation 25 February 2014, February 2014. URL: `http://www.w3.org/TR/turtle/`.

2 Julia Bosque-Gil, Jorge Gracia, and Elena Montiel-Ponsoda. Towards a Module for Lexicography in OntoLex. In *Proc. of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets at 1st Language Data and Knowledge conference (LDK 2017), Galway, Ireland*, volume 1899, pages 74–84. CEUR-WS, June 2017. URL: `http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_5.pdf`.

3 Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, and Guadalupe Aguado-de Cea. Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case. In Ilan Kernerman, Iztok Kosem, Simon Krek, and Lars Trap-Jensen, editors, *Proc. of GLOBALEX'16 workshop at LREC'15, Portoroz, Slovenia*. European Language Resources Association, May 2016. URL: `http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf`.

4 Julie Coleman, editor. *Love, Sex, and Marriage: A Historical Thesaurus*. Number 118 in Costerus New Series. Rodopi, Amsterdam, 1999.

5 Thierry Declerck, Eveline Wand-Vogt, and Karlheinz Mörth. Towards a Pan European Lexicography by Means of Linked (Open) Data. In Iztok Kosem, Milo¨ Jakubíček, Jelena Kallas, and Simon Krek, editors, *Proceedings of eLex 2015*, Ljubljana/Brighton, August 2015. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., Trojina, Institute for Applied Slovene Studies.

6 Felipe Faria. Georges Cuvier et le premier paradigme de la paléontologie. *Revue de Paléobiologie, Genève*, 32(2):297–302, 2013. URL: `http://www.academia.edu/download/40773384/E15-Revue_Paleobiol-322-Malvesy_et_al._eds-G._Cuvier_IV.pdf#page=4`.

7 R. R. K. Hartmann. Thesauruses. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, pages 668–676. Elsevier, Oxford, second edition edition, 2006. URL: `http://www.sciencedirect.com/science/article/pii/B0080448542004545`.

8 Werner Hüllen. *A history of Roget's Thesaurus: Origins, development, and design*. Oxford University Press, Oxford, 2004.

9 Christian Kay and Marc Alexander. Diachronic and synchronic thesauruses. In Philip Durkin, editor, *The Oxford handbook of lexicography*, pages 367–380. Oxford University Press, Oxford, 2016.

10 Christian Kay, Jane Roberts, Michael Samuels, and Irené Wotherspoon, editors. *Historical thesaurus of the Oxford English Dictionary: with additional material from "A thesaurus of Old English"*. Oxford University Press, Oxford, 2009.

11 Christian Kay, Jane Roberts, Michael Samuels, and Irené Wotherspoon, editors. *Historical thesaurus of the Oxford English Dictionary*. Oxford University Press, 2010. URL: `http://oed.com/thesaurus`.

12 Fahad Khan, Javier E. Díaz-Vera, and Monica Monachini. Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web ? In *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference (ESWC 2016), Heraklion, Greece, May 30th, 2016.*, pages 37–46, 2016. URL: `http://ceur-ws.org/Vol-1595/paper4.pdf`.

13 Bettina Klimek and Martin Brümmer. Enhancing lexicography with semantic language databases. *Kernerman DICTIONARY News*, 2015.

14 Lexicon Model for Ontologies: Community report, 10 May 2016, May 2016. URL: `http://www.w3.org/2016/05/ontolex/`.

15 Iseabail Macleod, Pauline Cairns, Caroline Macafee, and Ruth Martin, editors. *The Scots thesaurus*. Aberdeen University Press, Aberdeen, 1990.

16 M. Lynne Murphy. Meaning relations in dictionaries: Hyponymy, meronymy, synonymy, antonymy, and contrast. In Philip Durkin, editor, *The Oxford handbook of lexicography*, pages 439–456. Oxford University Press, Oxford, 2016.

17 J. Roberts, C. Kay, and L. Grundy. *A Thesaurus of Old English: In Two Volumes*. Rodopi, 2000.

18 Peter Mark Roget. Thesaurus of English Words and Phrases, Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition, 1852.

19 SKOS Simple Knowledge Organization System Reference: W3C recommendation 18 August 2009, August 2009. URL: `http://www.w3.org/TR/skos-reference/`.

20 Marvin Spevack. *A Shakespeare thesaurus*. Georg Olms AG, Hildesheim, 1990.

21 Sander Stolk. OntoLex and Onomasiological Ordering: Supporting Topical Thesauri. In *Proc. of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets at 1st Language Data and Knowledge conference (LDK 2017), Galway, Ireland*, volume 1899, pages 60–67. CEUR-WS, June 2017. URL: `http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_3.pdf`.

22 XKOS: An skos extension for representing statistical classifications, January 2017. URL: `http://rdf-vocabulary.ddialliance.org/xkos`.

23 Katharina Zeppezauer-Wachauer. Mittelhochdeutsche Begriffsdatenbank, 1992-2018. URL: `http://www.mhdbdb.sbg.ac.at/`.

PART III

Disseminating and Evaluating *A Thesaurus of Old English* as Linguistic Linked Data

Chapter 6

Information:
The paper that starts on the next page was published in 2019, available open access under the CC BY-SA license. The only change to that paper, here, is the inclusion of two numbers for the benefit of readers of the dissertation: the overall page number and the chapter number (presented in the margin in a grey box and a white box, respectively). When citing, please refer to the original publication and its page numbering.

Publication:
Sander Stolk, '*A Thesaurus of Old English* as Linguistic Linked Data: Using OntoLex, SKOS and *lemon-tree* for Bringing Topical Thesauri to the Semantic Web', Proceedings of the eLex 2019 conference, Sintra, 1-3 October 2019, pp. 223–247. http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_3.pdf.

A Thesaurus of Old English as Linguistic Linked Data: Using OntoLex, SKOS and *lemon-tree* to Bring Topical Thesauri to the Semantic Web

Sander Stolk

Leiden University, Leiden, the Netherlands
E-mail: s.s.stolk@hum.leidenuniv.nl

Abstract

An increasing number of dictionaries are represented on the Web in the form of linguistic linked data, utilizing OntoLex-Lemon for this purpose. Lexicographic resources other than dictionaries, however, have thus far not been the main focus of efforts surrounding this model. In this paper, we discuss porting a topical thesaurus to the Web: *A Thesaurus of Old English*. By means of this case study, this paper discusses how this thesaurus – and topical thesauri in general – can be represented with OntoLex-Lemon, SKOS and *lemon-tree* through a fully automated process. Along with discussing the terminology required for expressing *A Thesaurus of Old English* as linguistic linked data, this paper indicates challenges encountered in the conversion process. These challenges range from material that is not meant to be made available to the general public to distinctions and relations that have been left implicit in the legacy form but are of much value and, indeed, required to be expressed explicitly in its linked data form. The aim of this paper, thus, is to provide recommendations for representing topical thesauri on the Web and to grant insight into aspects that may be encountered in porting similar lexicographic resources in the future.

Keywords: thesaurus; linguistic linked data; conversion; automation

1. Introduction

An increasing number of dictionaries are represented on the Web in the form of linguistic linked data using the OntoLex-Lemon vocabulary (Bosque-Gil et al., 2016; Khan, 2016). Such a representation is thought to facilitate interoperability across linguistic resources, have the potential to increase their visibility, and promote their reuse (Declerck et al., 2015; Klimek & Brümmer 2015). However, lexicographic resources other than dictionaries have thus far not been the main focus of efforts surrounding OntoLex-Lemon and its modules. In this paper, we discuss porting a topical thesaurus to the Semantic Web: *A Thesaurus of Old English.*

A Thesaurus of Old English captures the lexis of the early medieval variant of English, spoken between roughly 500 and 1100 by the Anglo-Saxons (Roberts et al., 2015). This lexicographic resource presents a feature common to topical thesauri but uncommon to dictionaries: its topical system (i.e., a hierarchy of categories) that organizes lexical senses according to their meaning (Kay & Alexander, 2016). Moreover, this thesaurus

also distinguishes conceptual levels within the topical system – a feature that was already present in the first modern thesaurus, *Roget's Thesaurus* (1852). By means of this case study, then, this paper presents areas problematic for representing *A Thesaurus of Old English* – and topical thesauri in general – in OntoLex-Lemon alone, and turns to the novel model *lemon-tree* for the needed expressivity. This model combines OntoLex-Lemon with the SKOS vocabulary, filling minor but important lacunae perceived for topical thesauri specifically, thereby increasing the portability and interoperability of these lexicographic resources (Stolk, 2019).

Next to treating the terminology required for porting *A Thesaurus of Old English* to a linguistic linked data form, this paper will indicate further challenges in this process. These range from material available in the legacy form that is not meant to be made available to the general public (e.g., notes purely editorial in nature) to distinctions and relations that have been left implicit in the legacy form but are of much value and, indeed, required to be expressed explicitly in its linked data form. The aim for this paper, thus, is to provide recommendations for representing topical thesauri on the Web and to grant insight into aspects that may be encountered in porting similar lexicographic resources in the future.

2. *A Thesaurus of Old English*

A Thesaurus of Old English (TOE) captures the lexis of Old English. The words and their senses of this historical variant of English, spoken roughly between 500 and 1100, are grouped together in sets of synonyms and placed in an overarching hierarchy of categories. In addition, TOE indicates the distribution of words in the surviving Old English texts. Thus, some are flagged as found only in poetic works or as glosses. As of May 2017, the thesaurus contains 51,483 senses that have been sorted and categorized manually in 22,451 categories[1]. Accumulating and editing this wealth of information for the first publication of the thesaurus in 1995 took a team of scholars – led by Christian Kay, Jane Roberts, and Lynne Grundy – over fifteen years (Roberts, 1978). The fruit of their labour has certainly not gone unnoticed in the scholarly field concerning Old English.

Since its publication, TOE has been met with high praise. Rolf Bremmer Jr, for instance, states that the thesaurus fills a "voluminous gap [...] on the shelf of lexicographical tools" available for Old English (2002). Richard Dance, too, calls TOE "invaluable" for lexical studies and deems it an "impressive piece of scholarship" (1997). Manfred Görlach goes so far as to state that TOE is "the most important contribution to Old English studies for years", as its content allows scholars to "investigate what distinctions Anglo-Saxons felt important enough to make in the lexicon" (1998). This historical thesaurus, then, is considered a valuable asset to many scholars. Opening up

[1] These numbers are based on an export of the TOE database provided on 26 May 2017.

the knowledge contained within – by providing the thesaurus in an appropriate form – is therefore an important aspect for its use in research.

Work on TOE continued after its first publication in 1995, resulting in further editions. None of these, however, was published in a linguistic linked data form. The benefits promised by such a form – e.g., interoperability and reuse – warrants looking into how such a lexicographic resource can be represented using the relevant standards. This paper therefore details the process of bringing TOE to the Semantic Web. This process, which converts the contents of the current TOE database into the desired linked data form is illustrated with *frēols* (in the sense of 'free, not enslaved', see DOE, s.v. 'frēols adj.') that is positioned in the TOE category "Freedom, being free". This lexical sense and the category it belongs to are depicted in Figure 1 along with relevant context in the form of synonymous senses (cf. *frēot*) and superordinate categories from the topical system.

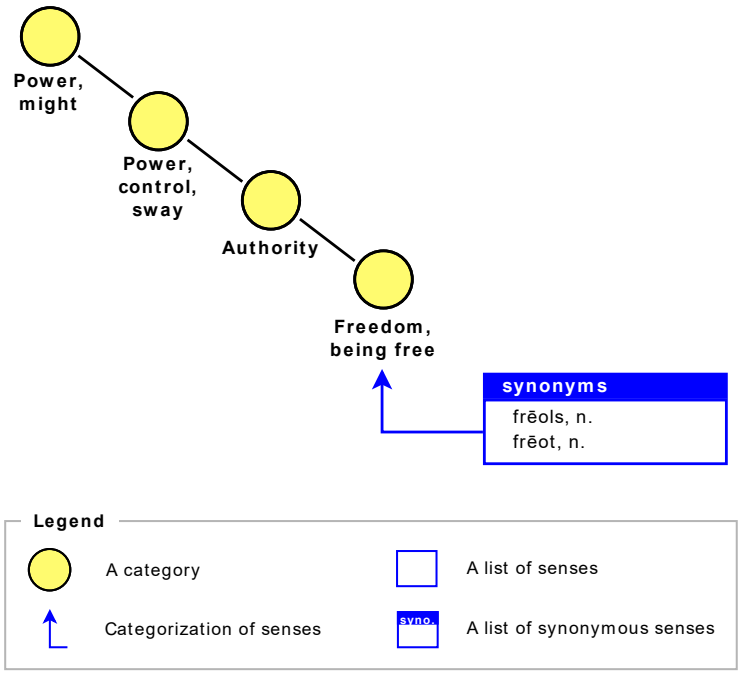


Figure 1: Sample of content from TOE.

In order to discuss the conversion process, we will first continue to describe the current digital form of the TOE database, referred to as its legacy form. The subsequent section provides a better insight into the desired, linguistic linked data form of TOE, which leverages the compact *lemon-tree* model for topical thesauri (Stolk, 2019) alongside the W3C standards OntoLex-Lemon and SKOS (*OntoLex*; *SKOS*). Finally, the conversion

process itself between these two forms is described, followed by the conclusion.

3. Legacy Form

The electronic edition of TOE hosted by the University of Glasgow employs a MySQL database to retrieve and display the thesaurus contents in webpages (TOE, 'Creation of the *Thesaurus*'). The database format is a tabular one, which makes exports possible to other formats that can capture rows and columns (MySQL 5.7 Reference Manual, 'What is MySQL?'). Such formats include Excel spreadsheets and CSV files (MySQL 5.7 Reference Manual, 'Alternative Storage Engines'). In fact, the University of Glasgow provides licensees of the TOE database with a copy by means of such formats. The version of the database provided for this research dates from 26 May 2017.

The TOE database consists of three tables. Each of the tables start with a single row containing the column headings. The rows below it – also known as records – capture instances. The first table discussed here is the category table of TOE, of which the structure is illustrated by Table 1.

catid	**t1**	**t2**	**t3**	**t4**	**t5**	**t6**	**t7**	**subcat**	**pos**	**heading**	**notes**
1	1								N	Earth, world	
2	1							1	N	As God's creation	xr Religion
3	1							1.01	N	In the beginning	
...	...	...	...	...	...	...	...	...	...	...	...
17187	12	1	1	9				18	V	To accept as a slave	
17188	12	1	1	9				19	V	To bring into bondage	
17189	12	1	1	10					N	Freedom, being free	
17190	12	1	1	10				1	N	Citizenship	
17191	12	1	1	10				2	N	A free man	
17192	12	1	1	10				3	N	A free woman	
17193	12	1	1	10				4	N	Freeman of lowest class	

Table 1: Structure of the TOE category table
(the category "Freedom, being free" is highlighted).

The category table of TOE is used to capture information on categories, where each record represents a single category. The table contains twelve columns in total:

- `catid`: This column acts as primary key, which "uniquely identifies each record in a database table" (*W3Schools.com*, 'SQL Primary Key').

- `t1` to `t7`: These columns capture the location in the taxonomy. Values in `t1` specify the position of the first main category compared to others at the same level, values in `t2` of the second tree level, and so on.

- `subcat`: This column indicates the location further down the taxonomy on a subcategory level (where applicable). Subcategories are distinguished from main TOE categories, which are indicated by `t1` through `t7`, in order to indicate a conceptual level in the taxonomy with smaller semantic differences than is the

case with main categories (TOE, 'Classification'). The subcategory position is not stored separately per subordination step, as the case with `t1` to `t7`, but as a single concatenated string delimited by stops.

- `pos`: This column stores the part of speech associated with a category. An indicated part of speech applies to all lexemes and their senses that are positioned directly at the category (i.e., they are not assigned to subordinate categories). Such a group of lexemes and senses in TOE always shares a single part of speech. Possible values are "aj" for adjective, "av" for adverb, "cj" for conjunction, "in" for interjection, "n" for noun, "p" for preposition, "ph" for phrase, "pn" for pronoun, "v" for verb, "vi" for intransitive verb, and "vt" for transitive verb (which may be monotransitive or ditransitive).

- `heading`: This column contains the name of each category in present-day English.

- `notes`: This column contains notes that are mostly editorial in nature. These include adjustments that have taken effect, matters still to be discussed, and so on. Due to their nature, the notes have so far been left unpublished in both paper and electronic editions.

Table 1 is identified by the key value 17189, called "Freedom, being free", expressed by nouns, and located in the taxonomy at position 12.01.01.10 – the 12th top category, followed by the 1st subordinate one, etc. Note that subordination relations applicable to given categories are not captured explicitly in this table but need to be deduced from the position in the taxonomy. Thus, the "Freedom, being free" category is understood to have the category located at 12.01.01 in the taxonomy as its direct superordinate category: "Authority" (`catid` 169410).

The TOE table discussed next is the category-xref table, of which a sample is shown in Table 2.

xid	catid	refid	tnum
1	18	588	01.03.01.05.01
2	18	9166	05.10.05.04.09
3	45	478	01.02.01.01.03
...	...	...	...
839	17189	16858	11.12.01
840	17189	18102	12.07.03

Table 2: Structure of TOE category-xref table
(the cross-references available at category "Freedom, being free" are highlighted).

Each record in the category-xref table represents a cross-reference in TOE from one category to another. Such a cross-reference indicates a related category that may be of interest to the user, too, but is found in another branch of the taxonomy. The table for

these cross-references contains four columns in total:

- `xid`: This column acts as primary key.
- `catid`: This column acts as foreign key. Such a key links one table to another by means of a reference to a primary key (*W3Schools.com*, 'SQL Foreign Key'). In this case, the column values refer to the primary key of the TOE category table. The categories indicated here are those at which a cross-reference is made.
- `refid`: This column, too, acts as foreign key to the TOE category table. The categories indicated here are those to which a cross-reference is made.
- `tnum`: The values of this column capture the location in the taxonomy of the category referenced in the `refid` column. (Note that this information is superfluous, as it can already be retrieved from the TOE categories table.)

To illustrate, the category "Freedom, being free" (`catid` 17189) has two cross-references: one to category "Absence of restraint, freedom" (`refid` 16858) and one to "Abstinence/exemption (from)" (`refid` 18102). These two categories referred to are found in another branch of the taxonomy than "Freedom, being free". In other words, there exists no subordinate/superordinate relation between them. Hence, the cross-referencing mechanism is employed to indicate that, nonetheless, these categories have a related topic according to the editors.

lid	catid	prefix	word	catorder	et	notes	oflag	pflag	gflag	qflag
1	1		brytengrundas	1		ChristA 355	Y	Y	N	N
2	1		brytenwangas	2		ChristA 380	Y	Y	N	N
3	1		eormengrund	3		Beo 859	Y	Y	N	N
...	...	...	...	...	...	...	...	...	...	...
39486	17187		hēafod niman	1			N	N	N	N
39487	17188	=	(ge)hæftan	1			N	N	N	N
39488	17189		frēols	1			N	N	N	N
39489	17189		frēot	2			N	N	N	N
39490	17190		burhrǣden	1			Y	N	Y	N
39491	17190		burhscipe	2			N	N	N	N
39492	17191		bonda	1	bond		N	N	N	N
39493	17191		ceorl	2	churl		N	N	N	N

Table 3: Structure of TOE lexeme table
(the lexeme frēols that is found at category "Freedom, being free" is highlighted)

From the data it appears that cross-references in TOE occur between main categories only. No cross-references exist from one subcategory to another, from a main category to a subcategory, or vice versa. Thus, although we find "Freedom, being free" is related to "Absence of restraint freedom", no cross-reference is made at one of its subcategories. It is likely that the editors of TOE deemed using cross-references for subcategories to be too fine-grained to indicate and maintain, and therefore kept such references confined to the main categories of the thesaurus. The third and last table of the TOE

legacy form is the lexeme table, depicted in Table 3.

Each record of the lexeme table represents an Old English lexeme that has been categorized based on one of its senses. The table contains eleven columns:

- `lid`: This column acts as primary key.
- `catid`: This column acts as foreign key to the TOE category table and assigns a lexeme, or rather one of the senses of a lexeme, to the category indicated.
- `prefix`: Values in this column, if filled in, can be "+" or "=". These signs correspond to + and ± in the second edition of the Old English dictionary by Clark Hall (CASD)[2]. Its introduction states the following:

  > Words beginning with *ge-* have been distributed among the letters of the alphabet which follow that prefix, and the sign + has been employed instead of *ge-* in order to make the break in alphabetical continuity as little apparent to the eye as possible. The sign ± has been used where a word occurs both with and without the prefix.

  This information on *ge-* prefixes has been superseded in TOE[3]. The current knowledge on prefix use can be deduced from the values in the `word` column.
- `word`: This column contains the head-form of each Old English lexeme. Optional segments of a word (which can be prefixes like *ge-*) are indicated between parentheses. See, for example, the lexeme with `lid` 39487 in Table 3.
- `catorder`: The values of this column indicate the order in which categorized lexemes are to be displayed that are located at the same category.
- `et`: This column contains etymological notes related to the lexeme. For instance, the Old English *ceorl* (`lid` 39493) developed into *churl* (OED, s.v. 'churl, n.').
- `notes`: This column contains notes. These typically mention how often or where a lexeme is found in the Old English corpus. Thus, the noun *eormengrund* (`lid` 3) is noted to be found on line 859 in the poem *Beowulf*.
- `oflag`: This column represents one of the distribution flags of TOE. When the value "Y" is recorded, the word form of the lexeme in question – not in any one specific sense – is marked as "very infrequent" in the Old English corpus.

[2] Information gained in personal correspondence with prof. Marc Alexander (6 August 2017).

[3] One example of knowledge in the `prefix` column being outdated is found with the lexeme with `lid` 582. The `prefix` column suggests the *ge-* prefix of this lexeme is mandatory (+), but the word column indicates that is no longer considered to be the case: "(ge)mȳþe".

- `pflag`: A distribution flag marking those word forms found only in poetry.
- `gflag`: A distribution flag marking those word forms found only in glosses.
- `qflag`: A flag marking word forms as "highly dubious" (TOE, 'Distribution Flags').

To illustrate, the lexeme *frēols* has a sense categorized as belonging to category 17189, "Freedom, being free" (see `lid` 39488). This lexical sense is meant to be displayed as the first one of this category, with the synonymous sense of *frēot* (`lid` 39489) as the second one. The word-forms of *frēols* are not marked as occurring very infrequently in the Old English corpus, in poetry only, in glosses only, or as questionable.

The lexeme table of TOE is rather inefficient for editorial purposes. Each record provides information for a lexeme (such as its head-form, and the distribution of its word forms) but also for a specific sense of that lexeme (such as its placement in the topical system). In fact, the `lid` value of each record is not unique per lexeme. Instead, it is unique per lexical sense. Information on a lexeme is therefore often recorded multiple times and in multiple locations – in a record for each of its senses. When a structure allows redundancy of information, consistency is more difficult to ensure. Contradictory statements are certainly present in the current dataset[4]. Such defects will not be magically mended by porting TOE to linguistic linked data. What the process will do, however, is make a clearer distinction between lexemes (or lexical entries) and lexical senses, which may improve detection of inconsistencies.

4. Linguistic Linked Data Form

A linguistic linked data form for topical thesauri should reuse standardized terminology in order to be interoperable. OntoLex-Lemon and SKOS are highly suitable to this end for capturing both lexical items and a hierarchy of concepts that represent the topical system of a thesaurus. Content from TOE can thus be published on the Web in a form that is machine-interpretable and understood in a wider community. Figure 2 charts, in a coarse manner, the relation between the content from the TOE sample and the linked data terminology from SKOS and OntoLex-Lemon. The relation `a` in this figure, and throughout this paper, is shorthand for `rdf:type` and can be read as "is a" or "is of type" (*RDF 1.1 Turtle*). As can be seen in Figure 2, a categorized lexeme corresponds with a `LexicalSense` in the ontolex module from OntoLex-Lemon. Similarly, a TOE category corresponds with a `LexicalConcept`. Thus, the Old English words *frēols* and *frēot* have lexical senses that lexicalize the concept "Freedom, being free". Superordination between concepts, such as between "Power, control, sway" and "Power,

[4] The noun *earfoþsīþ*, for instance, has two categorized senses in TOE (`lid` 22631 and 32588). Their registered `pflag` values contradict one another – "Y" and "N" respectively – even though both senses share their word forms and the distribution of these forms.

might", is indicated through the `broader` relation from SKOS. A more thorough list of linked data terminology and corresponding TOE content is available in Table 4. Most of the TOE table elements translate directly to linked data counterparts, although there are a few exceptions. These exceptions, discussed below, are taken into account in the linked data form that is proposed for the content of TOE.

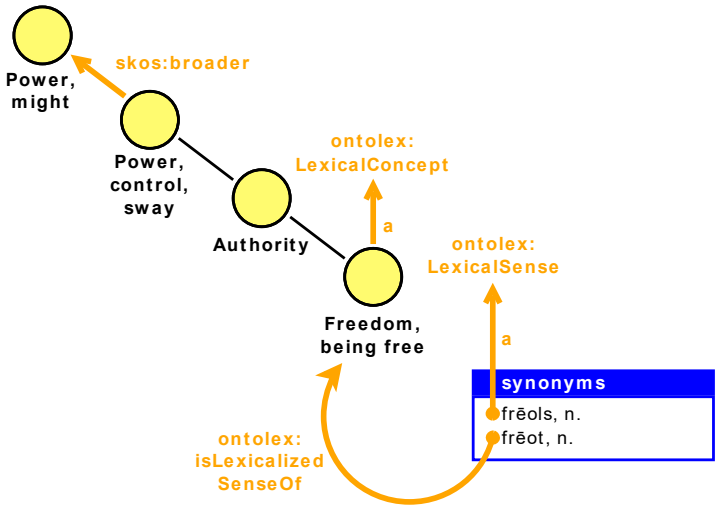


Figure 2: Sample of TOE content and its relation to linked data terminology from OntoLex-Lemon and SKOS

Firstly, some TOE content is not meant to be made available to the general public. Three elements are purely editorial in nature: the `notes` column from the category table and the `et` and `notes` columns from the lexeme table[5]. Various other elements are redundant or have been superseded. These bits may have been useful to the editors during the task of compiling the TOE dataset, but retaining them will likely prove detrimental or confusing. A case in point is the `catorder` column of the lexeme table. Although its values may aid in presenting synonymous senses in the desired order, they do not assist in determining the order for any given selection of senses. As the order of co-ordinate senses in TOE is a largely alphabetical one (with slight adjustments to take into account optional segments, length marks, and symbols specific to Old English), it would be possible – and preferable – to allow visualizations to determine the order of any selection of senses based on their head-forms. To this end, a label intended specifically for machines to order lexemes and their senses according to straightforward string comparison mechanisms (i.e., on ASCII characters only) would be easy to implement and utilize. The `prefix` column from the lexemes table, too, contains

[5] Information gained in personal correspondence with Prof. Jane Roberts (30 August 2017).

information that may best be left unshared with users. Its values are no longer current and can, especially if juxtaposed with the prefix information encoded in the `word` column, confuse users by contradictory statements on whether word forms of a particular lexeme existed with or without the *ge-* prefix. The aforementioned bits of information that are not meant for public consumption should not be part of any publication – including one in a linguistic linked data form.

Secondly, the TOE dataset is in some places more explicit than needed and less explicit in others. The category table, for instance, does not contain a column that explicitly captures the unique id (i.e., a `catid` value) of a superordinate category. As a result, subordination of categories needs to be deduced by means of combining the information from the identification columns – `t1` to `t7` and `subcat` – and comparing the identification values between categories. Storing the identification information separated over various columns hinders both retrieval of the identification string for a category and subsequent comparison of two such strings. Therefore, superordinate categories will be connected explicitly for the linguistic linked data form of TOE. Moreover, the identification string of each category will be stored and offered in a concatenated form rather than broken up in several segments[6].

Thirdly, the TOE dataset conflates information on lexical senses and lexemes into a single structure: the lexeme table. The linked data terminology from OntoLex-Lemon disentangles these two notions, calling the former a `LexicalSense` and the latter a `LexicalEntry`. As the primary key of the lexeme table is unique per sense of a lexeme, each of these records is associated with a `LexicalSense` rather than a `LexicalEntry`. Although the existence and name of a `LexicalEntry` can be deduced from the TOE lexeme table, the TOE dataset contains insufficient information to determine which senses belong to the same lexical entry. According to the specification of OntoLex-Lemon, words "may be different lexical entries if they are distinct in part-of-speech, gender, inflected forms or etymology" (OntoLex). Although TOE indicates the part of speech per lexical sense (i.e., via the `pos` column in the category table), the thesaurus does not currently indicate their gender or inflected forms. As such, a `LexicalEntry` will be created for each `LexicalSense` until information is made available in the future on which of these deduced lexical entries are meant to be one and the same. Such information can be compiled and offered by parties other than the editors of TOE, owing to the new linked data form of the dataset[7].

[6] The reason as to why the TOE category table does not store its identification information in a concatenated string but spread over multiple columns is likely found in the development process of the thesaurus, which saw shifts in the technologies used and the identification for categories (TOE, 'Creation of the *Thesaurus*'). One change in the identification system, for instance, is that subcategories have been provided with numbering since the first electronic edition.

[7] Asserting an `owl:sameAs` relation between two `ontolex:LexicalEntry` instances will effectively indicate that the two are to be considered one and the same entry.

Lastly, some of the contents of TOE require linked data terminology that is more specific than that found in SKOS and OntoLex-Lemon alone. To illustrate, a label used to aid computers in determining the presentation order of senses may be a `hiddenLabel` according to SKOS. Such hidden labels are intended for machine processing rather than for people to read. However, the hidden label for TOE should convey that it is specifically meant for the purpose of ordering rather than, for instance, searching alternative spellings. For this label, a new linked data term has been coined for TOE that extends the standardized terminology from SKOS. This coined term can be found in Table 4, including the terminology from SKOS that it extends (indicated through the '>' symbol). Next to this need specific to TOE, two other aspects of this thesaurus are in need of being captured in linked data – aspects shared by a great number of topical thesauri (Stolk, 2019).

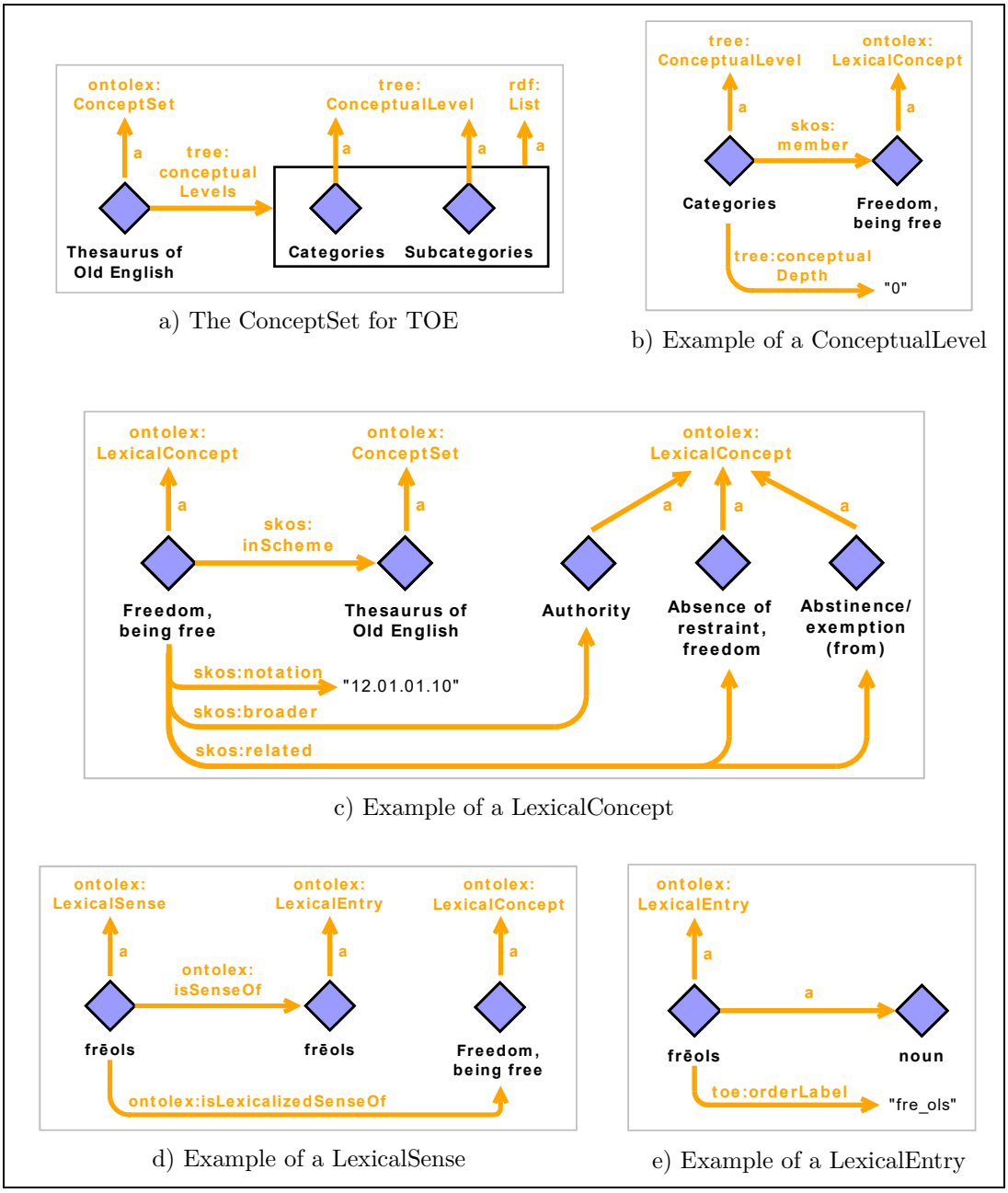
The first aspect common in topical thesauri is a division of their topical systems into conceptual levels. As mentioned above, TOE distinguishes two such levels in its database: main categories (simply called categories) and subcategories. The distinction of such levels has been deemed important enough to be included by editors. Indeed, for some thesauri, including TOE, the presentation and navigation mechanisms rely on these distinctions.[8] For a linked data form of TOE, then, this conversion follows the recommendations outlined by the compact *lemon-tree* model, which offers relevant terms such as `ConceptualLevel` and `conceptualDepth` – analogous to how tree levels can be represented using the XKOS (a well-known extension to SKOS used for statistics).

A second aspect, shared by all topical thesauri, is that they categorize lexical items. This is true both for thesauri that group lexical senses into sets of near-synonyms and those that do not. The *lemon-tree* model recognises the need to capture this loose form of categorization, for which it offers the `isSenseInConcept` property and indicates its relation to OntoLex terminology: the *lemon-tree* property is stated to be a more generic form (or super property) of OntoLex `isLexicalizedSenseOf`. This most basic form of categorization found in topical thesauri, then, can be automatically inferred by using the *lemon-tree* model alongside OntoLex for lexical senses in TOE that are asserted to lexicalize a given SKOS `Concept`. Figure 3 illustrates the resulting form for the sample content of TOE used throughout this paper. A combined presentation of this sample content is available in Figure 4. Prefixes are used to abbreviate the namespaces of data vocabularies, for which a mapping is provided in Table 5.

[8] Levels more abstract in nature are typically meant to be navigated first and allow the user to make greater semantic strides, as it were, than conceptual levels more specific in nature.

Linked data property	**Value obtained from legacy form TOE**
ontolex:ConceptSet	
skos:prefLabel	The name of the lexicon as a whole (i.e., "Thesaurus of Old English")
tree:conceptualLevels	An ordered list of the category types distinguished in the lexicon
skos:Collection > tree:ConceptualLevel	
skos:prefLabel	The name of the category type (i.e., "Categories" or "Subcategories")
tree:conceptualDepth	The conceptual depth of the category type
skos:member	The URI for a category belonging to this category type
ontolex:LexicalConcept	
skos:prefLabel	The name of the category
skos:broader	The URI for the superordinate category
skos:notation	The identification of the category
skos:related	The URI for a cross-referenced category
skos:inScheme	The URI for the lexicon as a whole (see ontolex:ConceptSet)
skos:topConceptOf	The URI for the lexicon as a whole (property applicable only to the top-most categories in the lexicon)
ontolex:LexicalEntry	
skos:prefLabel	The name of the lexeme
skos:hiddenLabel > toe:orderLabel	The name of the lexeme, rewritten so as to enable computers to sort these variants alphabetically by conventional means
rdf:type	The URI for the class indicating the part of speech of the lexeme
rdf:type	The URI for the class indicating the distribution of the word forms of the lexeme
ontolex:LexicalSense	
skos:prefLabel	The name of the categorized lexeme
ontolex: isLexicalizedSenseOf	The URI for the category at which the categorized lexeme has been positioned (and is therefore known to lexicalize)
ontolex:isSenseOf	The URI for the ontolex:LexicalEntry associated with the lexeme

Table 4: Linked data terminology and corresponding TOE content
(grey rows across the width of the table state the type of resource that will be formed; subsequent rows indicate which properties will be used to capture information for that resource and what their value will be).



a) The ConceptSet for TOE

b) Example of a ConceptualLevel

c) Example of a LexicalConcept

d) Example of a LexicalSense

e) Example of a LexicalEntry

Figure 3: Linguistic linked data form of TOE
(diamonds represent linguistic linked data resources of TOE; arrows represent properties).

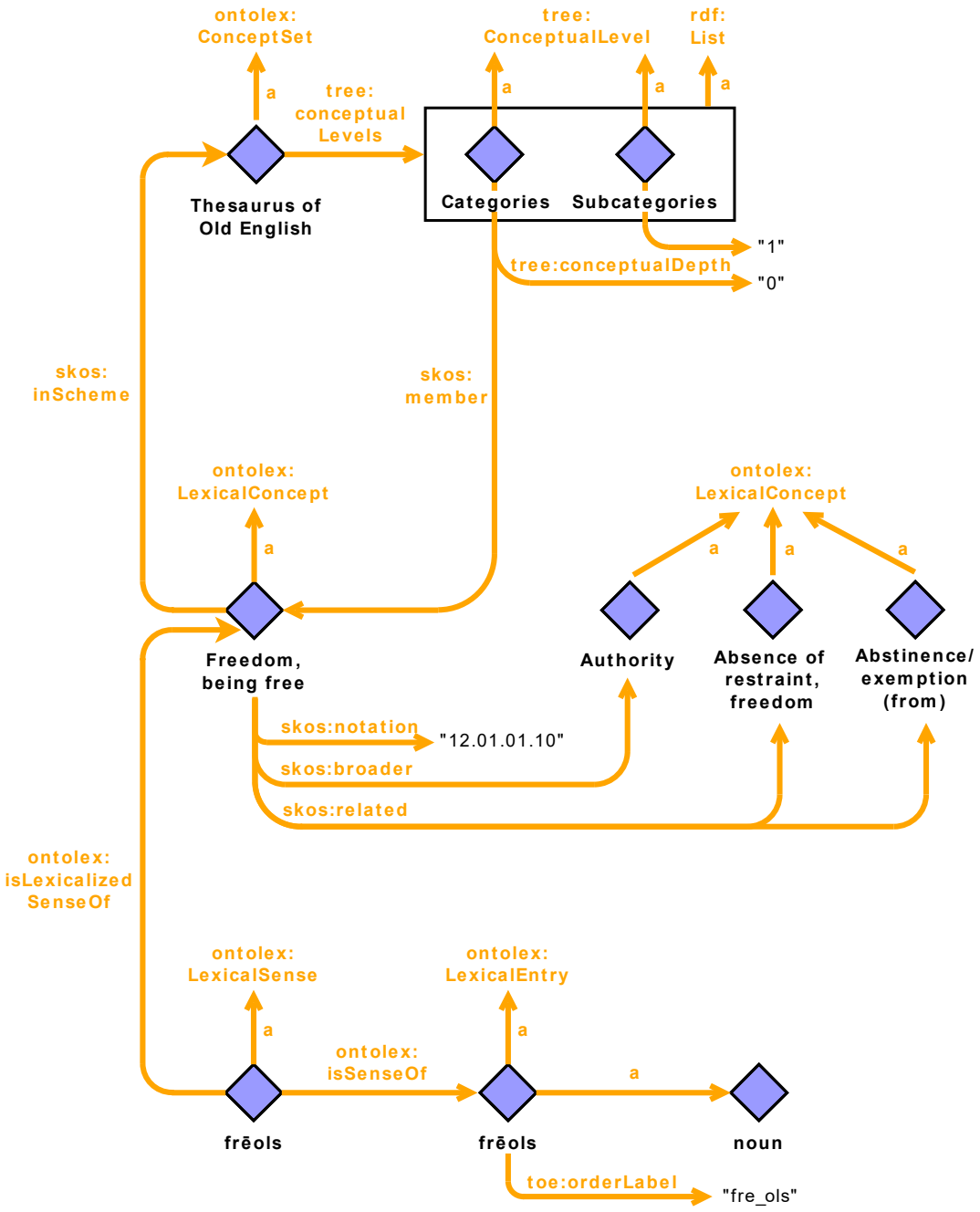


Figure 4: Linguistic linked data form of TOE
(combining the examples provided in Figure 3).

Prefix	Namespace
ontolex:	http://www.w3.org/ns/lemon/ontolex#
owl:	http://www.w3.org/2002/07/owl#
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs:	http://www.w3.org/2000/01/rdf-schema#
skos:	http://www.w3.org/2004/02/skos/core#
toe:	http://oldenglishthesaurus.arts.gla.ac.uk/
tree:	http://w3id.org/lemon-tree#

Table 5: Namespaces.

One further aspect needs to be discussed on bringing TOE content to the Semantic Web: the identification of each resource formed from TOE content. Bits of information on the Semantic Web are identified by a URI, typically in the form of an HTTP address. This holds for terminology from data vocabularies such as SKOS and OntoLex-Lemon, but also for instance data using such terminology. Best practices for coining URIs state that they should be simple, stable, and manageable (CoolURIs; CHIPS; SGOH). The first requirement entails that URIs need to be short and easy to remember; the second that they ought to be independent of the technology used to retrieve or visualize the content (as the software used may change); and the third that issuing new URIs should adhere to a straightforward strategy so as to be able to manage and maintain published content. With these requirements in mind, the following URI strategy has been adopted for the linguistic linked dataset of TOE. Each URI will be formed out of the following segments:

1. the Web domain of TOE (i.e., http://oldenglishthesaurus.arts.gla.ac.uk/),
2. the type of content the URI denotes (e.g., category, sense, entry), and
3. a unique number or string provided by the legacy form, if available.

The TOE category "Freedom, being free" (with `catid` 17189) thus gets the URI http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=17189 for its corresponding `LexicalConcept`. The lexical sense of the lexeme *frēols* (with `lid` 39488) gets http://oldenglishthesaurus.arts.gla.ac.uk/sense/#id=39488. This strategy has an additional advantage: it is aligned with the URI strategy in place for categories in the electronic edition of TOE hosted by the University of Glasgow. As a consequence, one can simply enter the URI of a category in a browser to view human-readable documentation on it. Adding linked data support to the electronic edition of TOE, as hosted by the University of Glasgow, is thus possible in the future without demanding a review or rework of the existing presentation. Having discussed both the original form of the TOE data and the desired linguistic linked data form, this paper will now turn to the conversion method employed to transform the former into the latter.

5. Conversion Process

Free digital tools already exist that facilitate a transformation from data in a tabular format to a linked data form. In selecting appropriate tools for the conversion of TOE from its legacy form to its desired linguistic linked data form, a number of requirements on the process need to be taken into account. These requirements, based on the premise that conversions ought to be reproducible by scholars with minimal effort, are listed in Table 6 and have been categorized according to priority[9]. Two requirements are mandatory, since these ensure an accurate conversion. The first is that the conversion process must accept tabular input either in an Excel spreadsheet or CSV format and provide transformed output in the RDF format (M1). The second requirement is that the process must be able to apply logic that relates the structure of the source to terminology from the desired linked data vocabularies (M2). The conversion logic for the TOE data has been described in Table 4. This logic also demands combining information from multiple tables, available in separate files. To illustrate, most of the information for lexical entries according to OntoLex-Lemon is found in the lexeme table of TOE. The part of speech of such an entry, however, is registered in another table of TOE: the category table.

Next to the requirements that are mandatory, three others have been formulated to which the process should adhere. Although not mandatory for an accurate outcome, these three requirements are geared towards increasing the maintainability and user-friendliness of the process. Firstly, the process should accept conversion logic in a form that has been standardized and is application-independent (S1). The alternative – relying on a format specific to a single tool – would limit the applicability, understandability, and reusability of the captured logic. Considering the availability of specific tooling and continued support from its creators are by no means guaranteed (as indeed seen for a number of conversion tools)[10], great reliance on a single tool should be avoided. Secondly, the process should be executable by scholars without a background in software development (S2). To be more specific, it should be possible to obtain and install the necessary tools without first having to compile the source code. Moreover, the tools should provide a visual user interface rather than only a command-line execution mechanism. Lastly, the conversion process should be automatable so that it can be performed again with minimal effort after an update of the thesaurus data (S3).

The final requirement for the process, assigned a lower priority than the foregoing ones, is meant to facilitate deploying and utilizing the resulting linguistic linked data. Web-based platforms will be able to retrieve and query information from a thesaurus if its

[9] The requirement prioritization follows the MoSCoW principles, developed by Dai Clegg et al. (1994).

[10] Availability and support for the tools AnnoCultor, Aperture, and NOR2O have been discontinued.

conversion output has been stored in a database that facilitates access for linked data technology (C1). A database for linked data content is called a triplestore. Triplestores typically allow accessing their stored content via queries using the standard querying language SPARQL, which web applications can use to interact with the data.

Must haves	
M1	Accept required input and output formats
M2	Apply required logic for conversion
Should haves	
S1	Employ standardized form for logic
S2	Allow for scholars to perform each step
S3	Allow for automation of all steps involved
Could haves	
C1	Store output in a triplestore with a query endpoint

Table 6: Requirements on the conversion process, categorized according to priority

The W3C provides a convenient overview of a number of tools that convert data into RDF (*ConverterToRdf*). Eighteen free tools listed there comply with requirement M1. These tools are listed in Table 7. Five of them appear to be discontinued, that is, they are no longer maintained or offered for download. Nine others do not comply with M2, either because they do not allow applying logic other than their default (Apache Any23) or because they cannot combine information from tables found in separate input files (RDF123; RDF Refine; csv2rdf4lod; Anzo for Excel; TabLinker; Excel2rdf; Sheet2RDF; Spread2RDF). The remaining four tools, then, conform to both mandatory requirements and should be able to convert the TOE legacy form into a linguistic linked data form. These tools are Datalift, Tarql, Virtuoso Sponger, and XLWrap.

One of the four remaining candidate tools for converting TOE data fails to meet requirement S1. This tool, XLWrap, defines its own form for capturing conversion logic, rather than using a standardized form (Langegger, 2017). A number of standardized forms for capturing conversion logic have been recommended by W3C. Two of these are specifically intended for logic converting tabular data into RDF: CSVW and R2RML. Unfortunately, these two forms are unsuitable for the conversion of TOE. The former cannot be used to combine information from multiple input files. The latter facilitates only relational databases as input and cannot be applied to Excel or CSV files. In fact, the three remaining tools – Datalift, Tarql, and Virutoso Sponger – facilitate transformations utilizing another logic form: SPARQL. This query language, standardized by W3C, allows selecting patterns from an RDF source and constructing new RDF data that adheres to desired patterns.

Software	**M1**	**M2**	**S1**	**S2**	**S3**	**C1**
AnnoCultor	*(discontinued)*					
Anzo for Excel	+	-				
Apache Any23	+	-				
Aperture	*(discontinued)*					
Convert2Rdf	*(discontinued)*					
csv2rdf4lod	+	-				
Datalift	+	+	+	+	-	+
Excel2rdf	+	-				
NOR2O	*(discontinued)*					
RDBToOnto	*(discontinued)*					
RDF Refine	+	-				
RDF123	+	-				
Sheet2RDF	+	-				
Spread2RDF	+	-				
TabLinker	+	-				
Tarql	+	+	+	-	+	-
Virtuoso Sponger	+	+	+	-	+	+
XLWrap	+	+	-	-	+	-

Table 7: Software tools and the requirements they meet

The way in which SPARQL is used differs between Tarql on the one hand and Datalift and Virtuoso Sponger on the other. Tarql employs a unique approach by running SPARQL directly on CSV input rather than on RDF data. It does this by emulating patterns have been found based on the tabular input. Datalift and Virtuoso Sponger employ SPARQL in a two-step transformation. First, these tools apply a default, direct mapping to obtain RDF data that is "often more geared towards describing the structure of the data rather than the data itself" (Lefrancois et al, 2017)[11]. This RDF data can subsequently be transformed to RDF data that uses the desired data vocabularies. In this second step, SPARQL (the standard query language for RDF data) is used to select patterns from the RDF source and construct new RDF data that adheres to the desired patterns. Indeed, this two-step approach is one that can be performed by end-users (using tools such as Datalink) but can also be automated (using a direct mapping application and any triplestore that supports SPARQL queries).

[11]The alternative solution proposed by these authors, an extension to SPARQL, appears promising but has not been accepted yet as part of the SPARQL standard proper.

Moreover, this two-step approach is also applicable to formats other than CSV, which may well suit future conversions beyond TOE. The conversion process for TOE, then, will employ the following generic steps:

1. obtain an RDF graph that expresses the structure of the input data
2. store the RDF graph in a triplestore
3. obtain the RDF that adheres to the desired linguistic linked data form through SPARQL queries

Taking these steps will also ensure that the last of the requirements, C1, is met. In other words, the desired linguistic linked data form that has been obtained will be available for queries by platforms that intend to visualize or utilize the thesaurus information. In fact, these three generic steps, here applied to TOE data, should be applicable to the conversion of any topical thesaurus, including those with legacy formats other than tabular data.

For the tabular data of TOE, the first step of the conversion process can be performed by a number of tools. Apache Any23, CSVW implementations[12], Datalift, and Apache Jena all express the structure of such input data in a similar manner. The default logic that these tools share when processing a CSV file is as follows. Firstly, these tools create a node in RDF for each record from the input. Secondly, they add a relation to that node for each of the filled in cell values they encounter. The identification of this relation (i.e., its URI) ends in the column name[13]. An example snippet of such output can be found in Listing 1. To obtain such results using Jena, one simply has to install Apache Jena and run the following command (adjusted to the desired input filename and the output filename):

```
> riot "input.csv" > "output-graph.ttl"
```

[12] See the CSVW report for a list of implementations (*CSVW Reports*).

[13] The initial letter of the column name is capitalized in the case of Apache Any23.

```
_:S39488 <file://C/lexemes.csv#lid> "39488" ;
<file://C/lexemes.csv#catid> "17189" ;
<file://C/lexemes.csv#word> "frēols" ;
<file://C/lexemes.csv#catorder> "1" ;
<file://C/lexemes.csv#oflag> "N" ;
<file://C/lexemes.csv#pflag> "N" ;
<file://C/lexemes.csv#gflag> "N" ;
<file://C/lexemes.csv#qflag> "N" ;
.
```

Listing 1: Snippet of RDF generated in the first step of the conversion process, based on the record for one of the senses of frēols (lid 39488) and expressed in the Turtle syntax.

The second and third steps of the conversion process require a triplestore. For this paper, the RDF4J triplestore is used to illustrate these steps. RDF4J offers a web-based interface, which allows users to set up a new repository for RDF content (see Figure 5) and therein store the intermediate RDF graphs obtained in step 1 (see Figure 6). Each of the graphs is assigned its own context in the repository, which will allow queries in the next step to select content accurately. Table 8 specifies the contexts used in the conversion process.

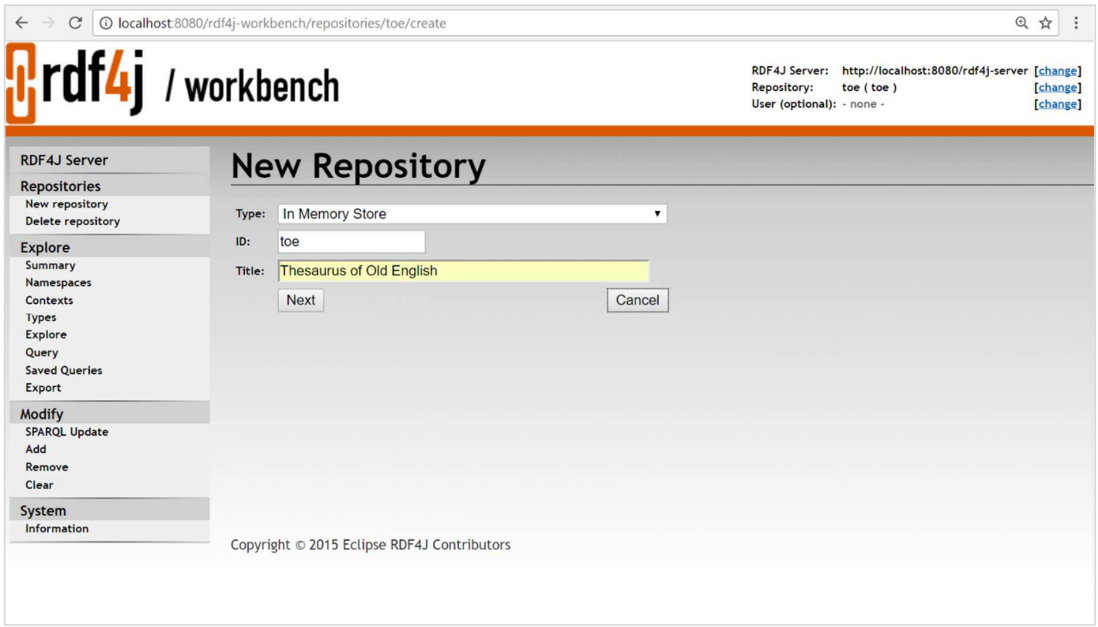


Figure 5: Creating a repository for TOE using the RDF4J user interface

Table of origin	Context
TOE category	<urn:toe:input:category>
TOE category-xref	<urn:toe:input:category-xref>
TOE lexeme	<urn:toe:input:lexeme>

Table 8: Contexts used upon adding RDF to the triplestore.

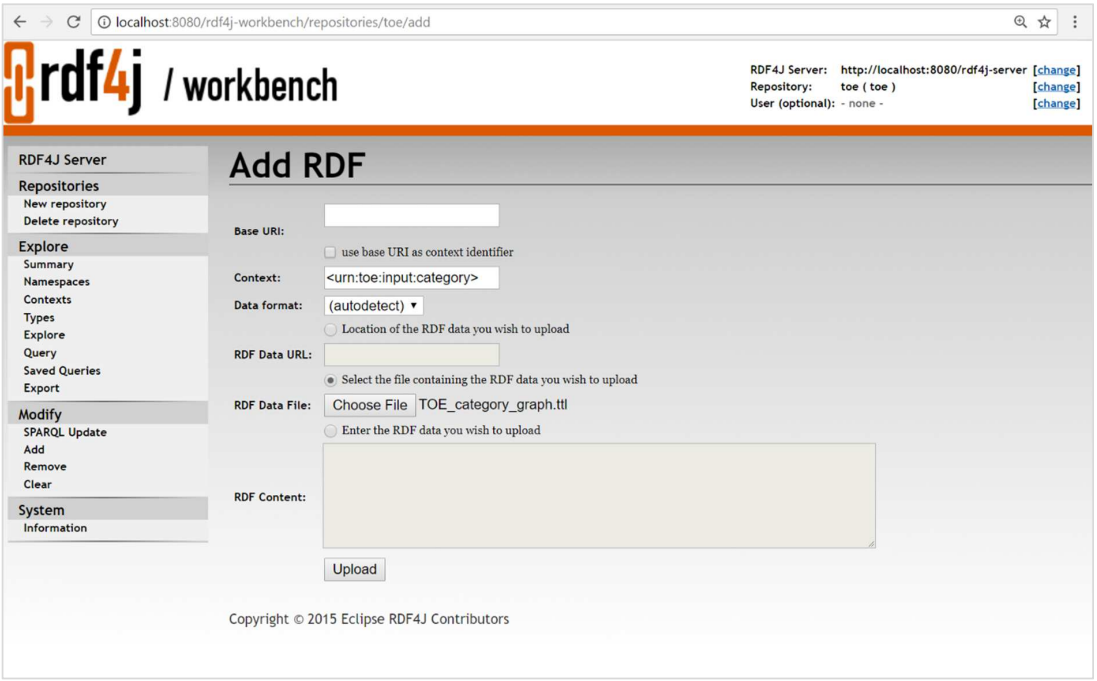


Figure 6: Adding RDF data to TOE categories using the RDF4J user interface.

In the third conversion step, queries are used to transform the available content in the repository to the desired linguistic linked data form. Such queries, written in SPARQL, can be executed via the RDF4J user interface (see Figure 7). Each query specifies a specific pattern that needs to be matched in the available content (in the WHERE clause of the query) and specifies another pattern that should be added as a result for each match (in the INSERT clause). Thus, patterns from the graph content of TOE can be transformed to patterns that conform to the desired outcome.

After the conversion, the resulting RDF will be available for querying and visualization. The intermediate RDF graphs that are uploaded in step 2 can be removed from the triplestore in order to ensure that only the final, desired form of the TOE dataset is indeed available in the repository. Automating the entire conversion process is also possible by means of a batch file. Both the batch file and queries that have been

employed in the conversion of TOE have been made available on GitHub[14].

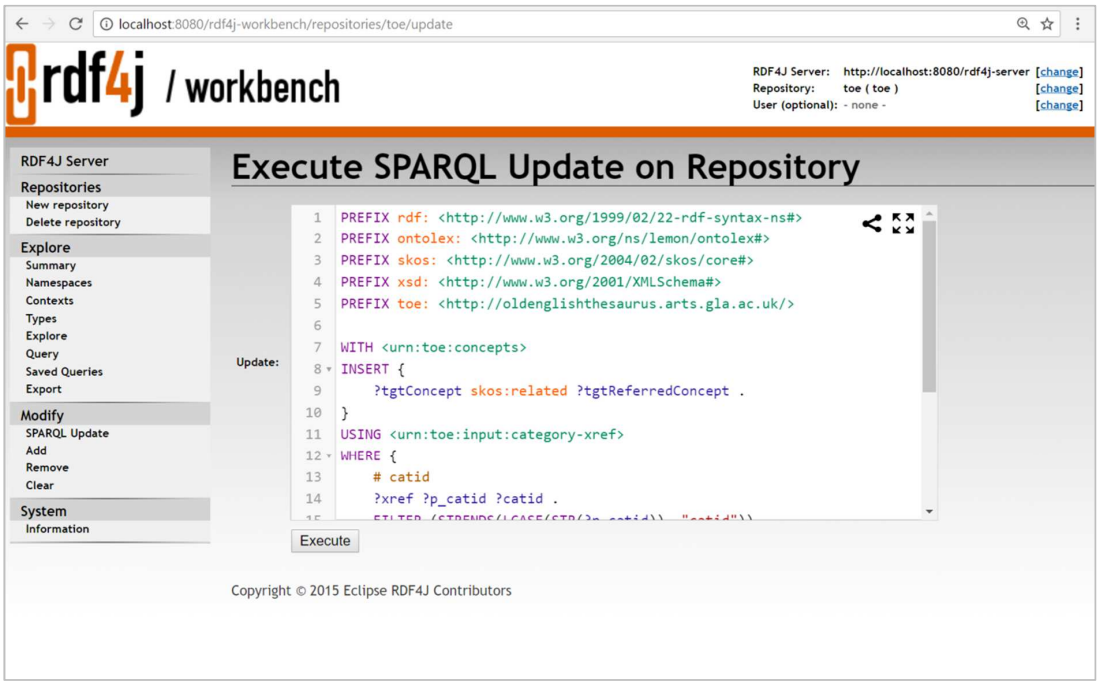


Figure 7: Executing a SPARQL update query using the RDF4J user interface

6. Conclusion

This paper has discussed the conversion of *A Thesaurus of Old English* from its legacy form to a linguistic linked data form utilizing OntoLex-Lemon, SKOS and *lemon-tree*. This conversion follows three steps: 1) obtaining an RDF graph that expresses the structure of the input data, 2) storing the graph in a triplestore, and 3) executing transformation logic using the standardized SPARQL language to produce the desired linguistic linked data form. Using SPARQL for capturing logic rather than a tooling-specific format ensures that the conversion process outlined does not rely on the existence of a single tool. Moreover, the three generic steps of the conversion process should be applicable to the conversion of any topical thesaurus – not just *A Thesaurus of Old English*. The results of the conversion discussed in this paper can be viewed in the online platform Evoke[15].

The new digital form of the thesaurus is used in a number of projects in order to investigate whether linked data mechanisms can facilitate research into Old English language and culture. Some of these projects link lexical items with information to

[14] https://github.com/ssstolk/lld/toe/

[15] http://evoke.ullet.net

indicate their presence in a specific Old English text. Thus, subthesauri can be fashioned to look into specific contexts. Other projects establish links between existing lexicographic resources – connecting ones on Old Dutch and Old Frisian with the thesaurus. Doing so allows for reuse of the thesaurus macrostructure for other languages, but also for contrasting the degree of lexicalization present in these historical languages (e.g., the number of words that we know to have been available in Old Frisian to express a given concept compared to that for Old English). The findings of these and further projects will be presented at the Exploring Anglo-Saxon Eloquence pre-conference workshop at the 21st International Conference of English Historical Linguistics[16].

7. Acknowledgements

The work described in this paper would not have been possible without the support of the Leiden University Centre for Digital Humanities for the Exploring Anglo-Saxon Eloquence project. Special thanks go out to the University of Glasgow, who have been kind enough to provide a license for working with the data of *A Thesaurus of Old English* and to give permission for distributing the resulting linked data form of the thesaurus on the Evoke platform.

8. References

AnnoCultor. Accessed at: https://sourceforge.net/projects/annocultor/. (4 June 2019)

Anzo for Excel. Accessed at: https://supportcenter.cambridgesemantics.com/docs/glossary/Anzo-Excel. (4 June 2019)

Apache Any23. Accessed at: https://any23.apache.org/. (4 June 2019)

Apache Jena. Accessed at: https://jena.apache.org/documentation/io/. (4 June 2019)

Aperture. Accessed at: http://aperture.sourceforge.net/. (4 June 2019)

Bosque-Gil, J. et al. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: The K Dictionaries Case. In I. Kernerman et al. (eds.) *Proceedings of GLOBALEX'16 workshop at LREC'16*. Portorož, Slovenia, pp. 65–72. Available at: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf.

Bremmer Jr, R. H. (2002). Treasure Digging in the Old English Lexicon, Review of *A Thesaurus of Old English*. *NOWELE,* 40, pp. 109–114.

CASD: *A Concise Anglo-Saxon Dictionary for the Use of Students*. (1916). 2nd edition. New York: Macmillan.

CHIPS: *Common HTTP Implementation Problems*. Accessed at: https://www.w3.org/TR/chips/. (9 June 2019)

Clegg, D. & Barker, R. (1994). *Case Method Fast-Track: A RAD Approach*. Boston:

[16] https://icehl21.wordpress.com

Addison-Wesley.

ConverterToRdf. Accessed at: https://www.w3.org/wiki/ConverterToRdf. (20 December 2017)

CoolURIs: *Cool URIs for the Semantic Web*. Accessed at: https://www.w3.org/TR/cooluris/. (9 June 2019)

CSV2RDF: *Generating RDF from Tabular Data on the Web*. Accessed at: http://www.w3.org/TR/csv2rdf/. (9 June 2019)

csv2rdf4lod. Accessed at: https://github.com/timrdf/csv2rdf4lod-automation/wiki. 4 June 2019)

CSVW Reports. Accessed at: https://w3c.github.io/csvw/tests/reports/. (9 June 2019)

Dance, R. (1997). Review of *A Thesaurus of Old English*. *Medium Ævum,* 66(2), pp. 312–313.

Datalift. Accessed at: https://datalift.org/. (4 June 2019)

Declerck, T. et al. (2015). Towards a Pan European Lexicography by Means of Linked (Open) Data. In I. Kosem et al. (eds.) *Proceedings of eLex 2015*. Sussex, United Kingdom, pp. 342–355. Available at: http://www.dfki.de/web/forschung/iwi/publikationen/renameFileForDownload?filename=eLex_2015_22_Declerck+etal.pdf&file_id=uploads_2536.

DOE: *Dictionary of Old English: A to I online*. Accessed at: http://www.doe.utoronto.ca. (4 June 2019)

Evoke. Accessed at: http://evoke.ullet.net

Excel2rdf. Accessed at: https://github.com/waqarini/excel2rdf. (4 June 2019)

Görlach, M. (1998). Review of *A Thesaurus of Old English*. *Anglia* 116(3), pp. 398–401.

Kay, C. & Alexander, M. (2016). Diachronic and Synchronic Thesauruses. *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 367–380.

Khan, F. (2016). Representing Polysemy and Diachronic Lexico-semantic Data on the Semantic Web. In *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference* (ESWC 2016). Heraklion, Greece, pp. 37–46.

Klimek, B. & Brümmer, M. (2015). Enhancing Lexicography with Semantic Language Databases. In *Kernerman DICTIONARY News,* 23.

Lefrancois, M. et al. (2017). A SPARQL Extension for Generating RDF from Heterogeneous Formats. In *Proceedings of the 14th International Conference of the European Semantic Web Conference*. Portorož, Slovenia, pp. 35–50.

Lemon-tree. Accessed at: https://w3id.org/lemon-tree. (4 June 2019)

MySQL 5.7 Reference Manual. https://dev.mysql.com/doc/refman/5.7/en/. (4 June 2019)

NOR2O. Accessed at: https://github.com/boricles/nor2o. (4 June 2019)

OED: *Oxford English Dictionary Online*. Accessed at: http://oed.com. (4 June 2019)

OntoLex-Lemon*: Lexicon Model for Ontologies*. Accessed at: http://www.w3.org/2016/05/ontolex/. (9 June 2019)

R2RML: *RDB to RDF Mapping Language*. Accessed at:

http://www.w3.org/TR/r2rml/. (9 June 2019)
RDF123. Accessed at: http://ebiquity.umbc.edu/project/html/id/82/RDF123. (20 December 2017
RDF Refine. Accessed at: http://refine.deri.ie/. (20 December 2017)
Roberts, J. (1978). Towards an Old English Thesaurus, *Poetica* 9, pp. 56–72.
Roget: *Thesaurus of English Words and Phrases, Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition*. (1852). London: Longman.
SGOH: *Style Guide for Online Hypertext*. Accessed at: https://www.w3.org/Provider/Style/URI. (4 June 2019)
Sheet2RDF. Accessed at: http://art.uniroma2.it/sheet2rdf/. (4 June 2019)
SKOS: *SKOS Simple Knowledge Organization Reference*. Accessed at: http://www.w3.org/TR/skos-reference/. (9 June 2019)
SPARQL*: SPARQL 1.1 Query Language*. Accessed at: http://www.w3.org/TR/sparql11-query/. (9 June 2019)
Spread2RDF. Accessed at: https://github.com/marcelotto/spread2rdf. (4 June 2019)
Spreadsheet-to-RDF Wrapper. Accessed at: http://xlwrap.sourceforge.net/. (4 June 2019)
Stolk, S. (2019). Lemon-tree: Representing Topical Thesauri on the Semantic Web. In M. Eskevich et al. (eds.) *Proceedings of the 2nd Conference on Language, Data and Knowledge (LDK 2019)*. Leipzig, Germany. Accessible at: https://doi.org/10.4230/OASIcs.LDK.2019.16.
TabLinker. Accessed at: https://github.com/Data2Semantics/TabLinker. (4 June 2019)
TOE: *A Thesaurus of Old English*. Accessed at: http://oldenglishthesaurus.arts.gla.ac.uk. (4 June 2019)
Turtle: *RDF 1.1 Turtle*. Accessed at: https://www.w3.org/TR/turtle/. (9 June 2019)
W3Schools.com. Accessed at: https://www.w3schools.com. (9 June 2019)
XKOS: *An SKOS Extension for Representing Statistical Classifications*. Accessed at: http://www.ddialliance.org/Specification/XKOS/1.0/OWL/xkos.html. (9 June 2019)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

http://creativecommons.org/licenses/by-sa/4.0/

Chapter 7

Information:
The article that starts on the next page was published in 2021, available open access under the CC BY license. The only change to that paper, here, is the inclusion of two numbers for the benefit of readers of the dissertation: the overall page number and the chapter number (presented in the margin in a grey box and a white box, respectively). When citing, please refer to the original publication and its page numbering.

Publication:
Sander Stolk, 'Evoke: Exploring and Extending *A Thesaurus of Old English* using a Linked Data Approach'. *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), pp. 318-358. doi: 10.1163/18756719-12340235.

BRILL

AMSTERDAMER BEITRÄGE ZUR
ÄLTEREN GERMANISTIK 81 (2021) 318–358

ABäG
brill.com/abag

Evoke

Exploring and Extending A Thesaurus of Old English *Using a Linked Data Approach*

Sander Stolk | ORCID: 0000-0003-2254-6613
Leiden University Centre for the Arts in Society, Leiden University, Leiden, Niederlande
s.s.stolk@hum.leidenuniv.nl

Abstract

This article provides an introduction to the web application Evoke. This application offers functionality to navigate, view, extend, and analyse thesaurus content. The thesauri that can be navigated in Evoke are expressed in Linguistic Linked Data, an interoperable data form that enables the extension of thesaurus content with custom labels and allows for the linking of thesaurus content to other digital resources. As such, Evoke is a powerful research tool that facilitates its users to perform novel cultural linguistic analyses over multiple sources. This article further demonstrates the potential of Evoke by discussing how *A Thesaurus of Old English* was made available in the application and how this has already been adopted in the field of Old English studies. Lastly, the author situates Evoke within a number of recent developments in the field of Digital Humanities and its applications for onomasiological research.

Keywords

Evoke – onomasiology – Old English – Digital Humanities – *A Thesaurus of Old English* – Linguistic Linked Data

1 Introduction

Vocabulary has been described as "a very sensitive index of the culture of a people" (Sapir, 1963: 27). This notion forms the corner stone of cultural linguistics, which explores the relationship between language and culture (Sharifian,

© SANDER STOLK, 2021 | DOI:10.1163/18756719-12340235
This is an open access article distributed under the terms of the CC BY 4.0 license.

2015: 515). Examining which words are (or were) available to a language community can offer valuable insights into their culture (Hough and Kay, 2017). Moreover, the word choices made *within* such a community can shed light on a number of aspects, including the intentions of speakers and authors, as well as their conscious or unconscious preferences.[1] Lexicographical works known as thesauri capture information about the lexicon and are veritable treasure troves for exploring the vocabulary of a language community and its relation to their culture.

A thesaurus is a lexicographic resource that organizes words and phrases according to their meaning rather than alphabetically.[2] Its overarching structure consists of a hierarchy of semantic concepts (Kay and Roberts, 1994). Concepts that represent abstract or generic meanings in this tree-like structure branch out to ones that are increasingly specific in meaning.[3] Each of these concepts can act as a category for words and phrases that express its meaning.[4] Thus, the semantic hierarchy of a thesaurus allows users to move from meaning to lexical item. This process can be illustrated with *A Thesaurus of Old English*, which allows users to look up early medieval English words by means of its semantic hierarchy (which it indicates through strings of numbers). An example of one of the most generic meanings in its hierarchy is the concept "12 Power, might". From this concept, it is possible to navigate branches down the hierarchy. One of these branches leads, via the more specific concepts "12.01 Power, control, sway" and "12.01.01 Authority", to the concept "12.01.01.10 Freedom, being free". Here, the thesaurus indicates which Old English words express this meaning: *frēols* and *frēot*.

It is self-evident that one of the main uses of thesauri is to look up alternative phrasings (e.g., Old English *frēols* or *frēot* to express 'freedom, being free').[5] However, owing to its semantic hierarchy, a thesaurus offers several other opportunities for research (Brewer, 2010: 802; Adamska-Sałaciak, 2010: 232; Busse, 2012: 88). These opportunities include exploring whether a word exists for a certain semantic concept; the number of words that express that concept (known as the degree of lexicalization or cultural elaboration) (Wierzbicka, 1997: 10–11); the number of words and nuances available within a given semantic

1 See also the contributions by Amos van Baalen and Thijs Porck in this special issue.

2 For further detail on this type of lexicographical work, and the distinction with other common senses of the word *thesaurus*, we refer the reader to Hartmann (2006) and Kay and Alexander (2016).

3 This structure, used to categorize lexical material, is not unlike the taxonomies of animals and plants created by Carl Linnaeus (1707–1778).

4 To be precise, thesauri categorize words and phrases in a given *sense* (Hüllen, 1999: 13).

5 See, for instance, the foreword to the first edition of Peter Mark Roget's thesaurus (Roget, 1852).

domain; which semantic domains are related through the various senses of a word (i.e., through polysemy); and how generic or specific the meaning of a word or group of words is. Traditionally, the acquisition of this kind of information from a thesaurus has relied on manual labour: leafing through paper editions of thesauri to find sections of relevance,[6] or manually counting the number of words presented on a website.[7] The web application Evoke (Stolk, 2018) introduces a user-friendly interface that allows researchers to digitally explore thesaurus content, enrich existing thesauri with additional knowledge, and analyse combined content in order to obtain semantic fingerprints that we call 'onomasiological profiles'.

The current version of Evoke (version 1.4.1) offers functionality to navigate, view, extend, and analyse thesaurus content.[8] The software facilitates users in answering such cultural linguistic questions as 'which words were available to a certain culture?', 'what is the degree of lexicalization of a concept' and 'which other notions were associated with a given word (owing to polysemy and homonymy)?' Moreover, thesauri that can be navigated in Evoke are expressed in an interoperable data form: Linguistic Linked Data (Cimiano et al., 2020). This form adheres to a number of best practices for data (Wilkinson et al., 2016; Lóscio et al., 2017) and employs standardized and interoperable data vocabularies for the Semantic Web (Miles and Bechhofer, 2009; Cimiano, McCrae and Buitelaar, 2016). In effect, this data form enables the extension of thesaurus content in Evoke with custom labels, facilitates linking thesaurus content to other digital resources, and allows novel analyses to be performed over multiple sources.

This article describes the development of Evoke and how it has been used to further explore *A Thesaurus of Old English*, which is the first thesaurus to be made available for analysis within this application. Sections 2 to 4 outline the requirements this web application had to meet in order to facilitate research into thesauri, its operating architecture, and the functionality it currently offers. Next, sections 5 and 6 discuss how *A Thesaurus of Old English* has been made available in a Linguistic Linked Data form and how researchers have been able to use Evoke to further explore this resource. Lastly, section 7 contrasts Evoke with existing software and discusses how Evoke and related work in the field of Digital Humanities may impact the future of onomasiological and cultural linguistic research.

6 Examples of thesauri available in print are Spevack (1993) and Macleod *et al.* (1990).

7 Examples of online thesauri include *TOE*, *HTE* and *HTS*.

8 The homepage of Evoke is http://evoke.ullet.net. The source code of the application is openly available at https://github.com/ssstolk/evoke/.

2 Requirements

In order to facilitate research into thesauri, the first step in making Evoke was to gauge the research needs that it could answer. On the basis of published reviews of thesauri as well as a number of research cases, five requirements were established that the new web application had to meet. The first three requirements (R1–R3) were gathered for the first version of Evoke. The remaining two requirements (R4–R5) were formulated for subsequent iterations of the application and were collected from stakeholders – experts in lexicography, linguistics and philology (amongst other fields) – to ensure that the software is intuitive and useful for both research and educational purposes. Stakeholder requirements were gathered by several means: dedicated stakeholder meetings, workshops, and feedback based on preliminary results in research and education projects. Additionally, requirements on the architecture (AR1–AR3) were based on best practices for data on the Web, transparent data management, and limitations imposed by licensing schemes of existing lexicographic resources.

R1. Navigation. A thesaurus can be approached in two manners: through its overarching taxonomy or through the lexis it organises (*HTOED*: ix). Both means are deemed key in allowing users – both newcomers and frequent users – to navigate the lexicographic content (Kay and Alexander, 2016: 368).

R2. Resource views. Complete overviews of available information are to be presented on any given resource within a thesaurus that a user chooses to inspect. These overviews must indicate relations to other resources where relevant, such as listing the words that are allocated to the viewed location of the thesaurus taxonomy, or indicating the various branches of the taxonomy that contain one of the senses of the currently inspected word or phrase.

R3. Extension. The application should allow users to extend a thesaurus, connecting additional information to its content (cf. Bremmer, 2002: 111; Görlach, 1998: 399; Dance, 1997: 313; Kay, 1996: 72). Examples of such extensions are indications of date and dialect, results from corpus searches, and indications whether a word or meaning is found in a particular text, context, or is notable in some other qualitative or quantitative way. This functionality offers users the means to have the thesaurus reflect their own interests and to share salient information with others.

R4. Analyses. Thesauri are valuable for investigations into a range of aspects encoded in the lexicon: cultural elaboration, semantic domains and their cultural connotations, stylistic preferences of authors, use and development of metaphors, and so on (e.g., Spevack, 1993; Crystal, 2014; Anderson, Bramwell, and Hough, 2016; Porck, 2016: 59–71, 239–294; Diller, 2017). Statistical analyses,

utilizing the onomasiological structure of the thesaurus and features of the lexis it contains, are therefore a key functionality.

R5. Data management. Proper data management should be an essential aspect of the application in facilitating onomasiological research. Users must have full control over their own data (e.g., creating backups, sharing their data with others) and the ability to select which data sources are deemed relevant for their explorations, allowing sets of information to be combined for viewing and analysis.

187

7

AR1. Interoperable data form. The software is to read and display resources published as Linguistic Linked Data. This vocabulary, and the underlying data format (RDF), facilitate reuse and interoperability of linguistic resources according to the FAIR data principles (Wilkinson et al., 2016; Lóscio et al., 2017).

AR2. Decentralized data. The software must be capable of accessing data stored in a decentralized manner rather than relying on a central database, thereby stimulating a separation of data storage and services. Such separation is intended to remove barriers in selecting alternative solutions on both fronts, i.e., where the data is hosted and which application suits the needs of the user best (Verborgh, Wrigley, and Ballardini, 2019).

AR3. Support limited licenses. It is not uncommon to find lexicographic resources subject to licenses meant for viewing only, stipulating that users are not allowed to copy or download a substantial portion of the entire resource.[9] *A Thesaurus of Old English* is one such work. The architecture of the application must allow interacting with and extending resources that are available under a license limiting access to browsing, in addition to those available under fully open access.

The eight requirements listed above (R1–R5 and AR1–AR3) have informed the design of Evoke, which will be discussed in the next two sections. These sections, on the architecture and functionality of the application, will reference requirements that are relevant for the design element under discussion.

3 Architecture of Evoke

Evoke has been designed as a web application, which, when accessed, runs in the user's internet browser. The application loads linguistic data from available data services and employs client-side rendering to display that information; rather than fetching an entire new webpage from a server whenever the user navigates to a different section, the application fetches only data necessary

9 See, for instance, *OED Online*, *HTE*, and the historical Dutch dictionaries part of the *Geïntegreerde Taalbank*.

(linguistic data, in this case) to fill out pages that it itself composes locally (Scott, 2015). Code for navigating Evoke is therefore executed on a user's computer rather than on a server dedicated to this purpose, resulting in a thin server architecture. As a result, Evoke demands server capabilities for hosting static files only (i.e., the web application) instead of additionally offering more advanced rendering technologies. The smaller demand on server-side resources should, in the case of Evoke, reduce hosting costs for the application. The code libraries used to render the interface of Evoke client-side are React and Reactstrap (basic HTML) complemented by ones specifically intended for vector graphics (i.e., D3, D3Pie, Recharts, and Wordcloud).[10]

3.1 *Linguistic Linked Data*

The data form supported by Evoke for accessing, exploring, and extending content is Linguistic Linked Data (cf. AR1) (Cimiano et al., 2020). This data form adheres to a number of best practices for data (Wilkinson et al., 2016; Lóscio et al., 2017) and employs standardized and interoperable data vocabularies for the Semantic Web (Miles and Bechhofer, 2009; Cimiano, McCrae, and Buitelaar, 2016). The Linguistic Linked Data paradigm builds on Linked Data principles, which advocate the use of Web mechanisms for capturing and sharing data, employing graph-based models (i.e., nodes and relations between them) and identifying data by means of IRIs (often Web addresses) (Cimiano et al., 2020). The use of IRIs allows one to capture and identify data in a language-independent manner, reuse terminology defined elsewhere, and to create links between datasets or nodes within different datasets. In effect, this data form enables thesaurus content in Evoke to be extended with custom labels and links to other digital resources.

Applying Linked Data principles to language resources nets a number of benefits (Chiarcos et al., 2013). One of these benefits is that their data form enables the merging of datasets in order to obtain a valid combined set of data. Thus, linguistic resources and datasets elaborating on them can be queried in unison. A second benefit is an increased level of interoperability. Using standardized terminology in describing linguistic data increases a shared understanding of that data and facilitates their interpretation by software. Moreover, the use of IRIs as identifiers ensures data can be linked without the need for duplication of information from one dataset into the other. The ability to link (or reference) in such a manner is valuable in the setting of Linguistic Linked Data, since it is not uncommon to find lexicographic resources subject

10 React: https://reactjs.org/. Reactstrap: https://reactstrap.github.io/. D3: https://d3js.org/. D3Pie: http://d3pie.org/. Recharts: https://recharts.org/. Wordcloud: https://wordcloud2-js.timdream.org/.

to licenses meant for viewing only, stipulating that users are not allowed to copy or download a substantial portion of their content (e.g., *OED Online* and *HTE*). By adopting IRIs in published lexicographic resources, their users can explore and extend these resources, engaging with the content offered, without infringing on such licenses (cf. AR3).

The Evoke web application, in order to draw on the aforementioned benefits, assumes language resources to adopt Linguistic Linked Data as specified by the W3C OntoLex community (Cimiano, McCrae, and Buitelaar, 2016), applied specifically to the context of topical thesauri (Stolk, 2019). Other resources can be viewed and extended in Evoke, too, as long as they are formulated using the same syntax that underlies Linguistic Linked Data: RDF (Cyganiak, Wood, and Lanthaler, 2014). Thus, Evoke can work with non-linguistic data connected to thesauri as well as future modules put forward by the OntoLex community for Linguistic Linked Data (such as that supporting frequency and attestations) (Chiarcos et al., 2020). The use of RDF in Evoke is not limited to solely the *content* of datasets available, as will be discussed in sections 3.3 and 3.4.

3.2 *Accessing Datasets*

The architecture of Evoke offers a high degree of flexibility in managing which content is to be viewed or analysed, including where individual datasets have been made available. The functionality to realize this hinges on the use of so-called data catalogues, which list available datasets and the data services supplying them. Upon start-up, Evoke loads a default data catalogue that, at the time of writing this article, contains the Linguistic Linked Data version of *A Thesaurus of Old English* (see section 5) and the datasets created through novel research presented in this special issue (e.g., *Beowulf Thesaurus*; *Andreas Thesaurus*; *Old English Martyrology Thesaurus*; *Ælfrician Vocabulary*; and *Old Frisian Kinship*). Custom data catalogues can, however, be used too.

Datasets listed in a catalogue are accessed through asynchronous calls to SPARQL endpoints and/or APIs (cf. the online repository in Figure 1). As data is retrieved through such calls in the background, the Evoke web application contains a so-called high-order component that acts as loader. This loader wraps the component that is to be presented, but still void of data, in another component that first awaits data requests. During the loading time, a loading icon is presented. Once the data required has been retrieved, the loader renders the wrapped component with its proper input. This mechanism is used, in combination with the underlying data form of Linguistic Linked Data, to follow links iteratively in fetching further information from all datasets that are to be accessed. Thus, when a user selects a semantic concept of a thesaurus for viewing, a list of words that express that concept is retrieved. For each of those words, their IRI (i.e., their identifier) is subsequently used to collect all

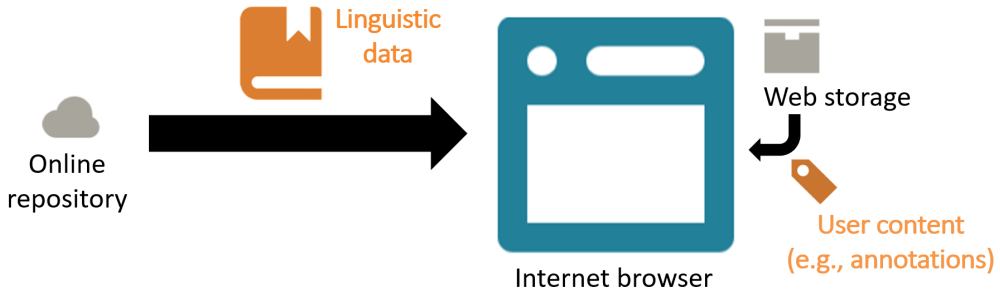


FIGURE 1 Sources of data that Evoke retrieves and presents

associated labels available across the datasets – whether they are part of the original thesaurus data or a set of annotations created by others.

The application reads, next to data served remotely via SPARQL or API calls, Linked Data stored in the user's internet browser (cf. the Web storage in Figure 1). A dataset in the browser can be stored using either the Turtle serialization or in JSON-LD and is interpreted using libraries from the Comunica framework (Taelman et al., 2018).[11] The use of this storage method in Evoke is detailed in section 4.3.

3.3 *Data Catalogue*

The data catalogues in Evoke adhere to the W3C DCAT vocabulary, an international standard developed specifically for expressing datasets and the services that provide these sets, including access details (Albertoni et al., 2020). This information is stored in the JSON-LD format and can therefore be read as JSON or, through the context provided, interpreted as RDF (Sporny et al., 2020). The use of these standards are meant to accommodate a higher level of interoperability with other tooling and services. An example catalogue is shown in Appendix 1. Drawing from a catalogue, the user interface of Evoke provides the means to select which available datasets are to be explored in unison (cf. R5, see Figure 9). Users can add their own datasets or data services to a catalogue, store their catalogue locally as a JSON file, and activate a local catalogue by dragging and dropping it onto the Evoke web application.[12]

Access to datasets may or may not need to be limited, depending on the usage license associated with them. In order to ensure that Evoke can work with more restrictive licenses (cf. AR3), two types of access mechanisms for data services are supported in data catalogues: (1) a SPARQL endpoint and

11 The Comunica library n3 is used for Turtle; jsonld-streaming-parser for JSON-LD; actor-init-sparql-rdfjs for SPARQL queries in browser memory.

12 Future plans include functionality to load a data catalogue from a URL, passed as a parameter to a deployed instance of Evoke, which will make it easier to explore different data catalogues in Evoke without 1) users loading these themselves or 2) administrators deploying separate instances of Evoke sporting different default catalogues.

(2) the Evoke API. The former allows services to respond to any query using the standardized querying language for RDF. The latter ensures content can be viewed and browsed in Evoke through a basic set of queries specific for this need, without offering users full access (that is, the possibility to extract or download the full dataset).[13] Distinguishing between these two types of access mechanism in the data catalogue is achieved through different values for the *endpointDescription* attribute of data services.[14]

Which datasets listed in the active catalogue will be queried by Evoke depends on the selection made by the user. The Evoke interface allows users to enable (or disable) datasets listed. Only those datasets can be enabled that (1) have a data service associated with them and (2) already have all of their required dependencies enabled. To illustrate, the 'Riddle 47' dataset contains links to the dataset 'A Thesaurus of Old English' and depends on it for analyses. Once a user has enabled this required dataset in the interface, that user can opt to enable the 'Riddle 47' dataset, too (see the top bars in Figure 9). Datasets served by the same service, though available in different graphs, are queried in unison and allow statistical analyses to be performed.[15]

3.4 *Browser Storage*

Any user of the Evoke web application can annotate linguistic content, such as words or semantic concepts, with information relevant to them. Typing a sentence in the annotation component of a page will automatically create a Linked Data annotation that adheres to the Web Annotation standard of W3C (Sanderson, Ciccarese, and Young, 2017), including any extracted label when a hashtag is used (see "#riddle47" in Figure 10). The novel aspect of this approach is that such an annotation is not stored in an online database, but is instead stored locally in the user's internet browser, employing the *localStorage* attribute of Web storage (cf. AR2) (Hickson, 2016). Annotations stored in

13 The Evoke API implements the queries Evoke relies on as a REST interface. Responses return SPARQL query results (see http://evoke.ullet.net/api). An open-source implementation of the API, which fetches results from a local SPARQL endpoint, is available at https://github.com/ssstolk/evoke-api/. Implementations of the API can, of course, also be built on top of technologies other than SPARQL endpoints (e.g., SQL).

14 The value of *endpointDescription* denoting a SPARQL endpoint is "http://www.w3.org/ns/sparql-service-description#Service"; the value denoting the Evoke API is "http://evoke.ullet.net/api".

15 Evoke does not perform statistical analyses over data found at *different* data services, since that would require assembling a full picture of the relevant data either at the user's internet browser, one of the data services, or a data aggregator (be they query results or parts of datasets), which is an intricate process, time-consuming, heavy on network traffic, and possibly restricted or made impossible by the licenses that are applicable to the datasets.

the browser can be downloaded as a file to backup (see Appendix 2) and can be reactivated in the browser – giving users full control over their created content and allowing them to share it in the manner of their choosing (cf. R5). Publishers of the original lexicographic resource benefit from this approach, too, as they neither need to moderate, store, or host annotations, nor offer users login mechanisms before they can interact with the information. Costs for hosts may thus be substantially reduced for presenting users with this functionality.

Annotations contain references to the identifiers, or IRIs, of the original lexicographic content without including the raw data of the annotated content in the annotation itself (cf. AR3). This approach allows users to already explore dictionaries and interact with them, formulate a plan of research, and at a later stage take the hurdle in getting support for further research from the publishers – be it in the form of a more open license, getting access to advanced services, or obtaining advice from lexicographers and other experts. Users may have an invested interest in the lexicographic resource at this point. Moreover, their additions are explicit, digital, and can be used in this form for analyses when queried in unison with the original dataset (facilitated by the characteristics of Linguistic Linked Data).

User data stored in the internet browser can, as with any RDF dataset, be published to a data service and added to a data catalogue for use in Evoke. In fact, when one publishes through the Evoke user interface, a new data catalogue is created automatically in which the published dataset is listed. This updated catalogue is not made public by the application: as with other user content, the catalogue is stored in the browser. A means to download the updated catalogue is provided to the user immediately after a successful publication. Users can choose to share it with others in a way they see fit or, if they wish to share the newly published dataset publicly, upload it to a public server and/or contact the administrator of the application to request inclusion of the dataset in the default catalogue of the deployed instance.

4 Features of Evoke

This section details the various features Evoke has to offer users in exploring *A Thesaurus of Old English* and other resources.

4.1 *Navigating*

As with print editions of thesauri, browsing the information within them is perhaps the single most fundamental need that users of digital editions have. The preface to the print edition of the *Historical Thesaurus of the Oxford*

English Dictionary states that there are "two ways to approach a thesaurus: by familiarizing oneself with its structure and principles of organization, or [...] by using its index to determine in which category or categories a word appears" (*HTOED*: ix). In the move from ink to Internet, these approaches have translated to navigating the semantic hierarchy of digital thesauri and to locating words using a search engine (e.g., *TOE* and *HTE*). The speed with which a user can locate a word in an electronic environment and access its information takes only "a twinkling of an eye" in comparison to print editions (Brewer, 2010: 804). The following subsections will discuss these two methods of navigation as implemented in Evoke – through the hierarchy and its search engine (cf. R1).

193

7

4.1.1 Navigating the Semantic Hierarchy

The semantic hierarchy of a thesaurus allows users to move from meaning to words or phrases that express this meaning (Hüllen, 2004: 282–283). Evoke presents such a hierarchy when opening a thesaurus. In the case of *A Thesaurus of Old English*, its 18 top concepts are shown (see Figure 2). Each of these concepts heads its own branch in the hierarchy, which can be navigated. Clicking on "Power, might", for instance, will navigate the branch headed by this semantic concept. The path of branches chosen is indicated through breadcrumbs: a trail of semantic concepts that represents the user's current location in the hierarchy (see Figure 3). These breadcrumbs can be used in navigation to

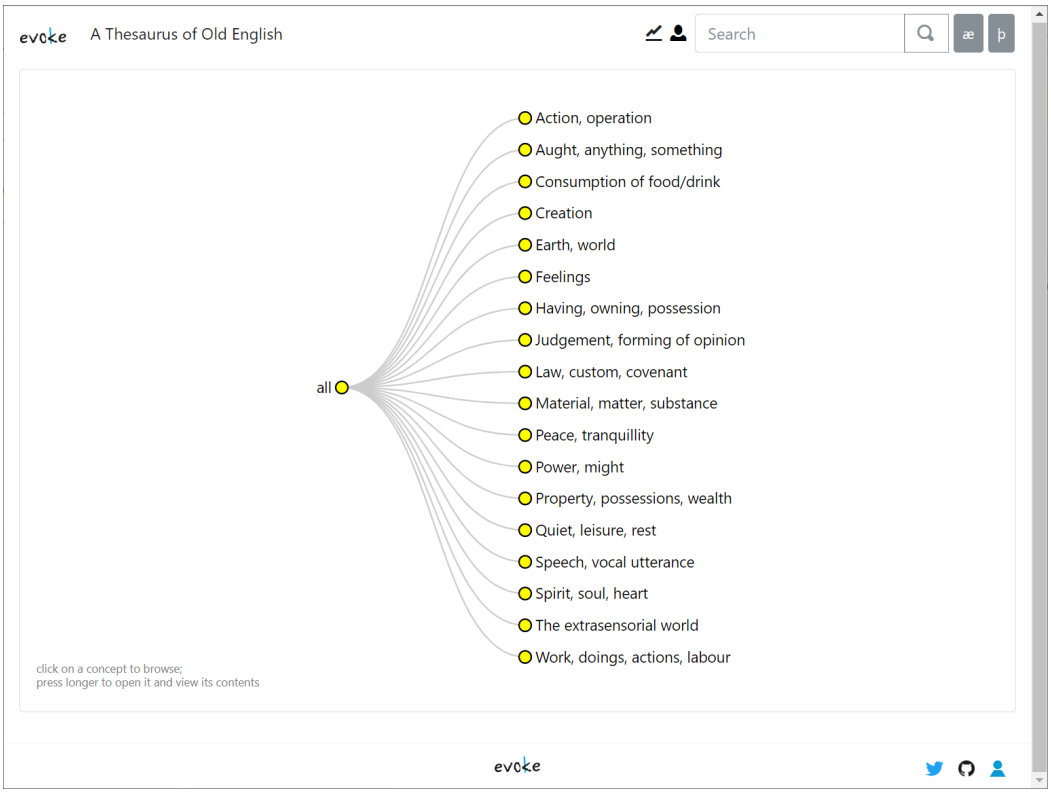


FIGURE 2 Navigating the top of the semantic hierarchy of *A Thesaurus of Old English*

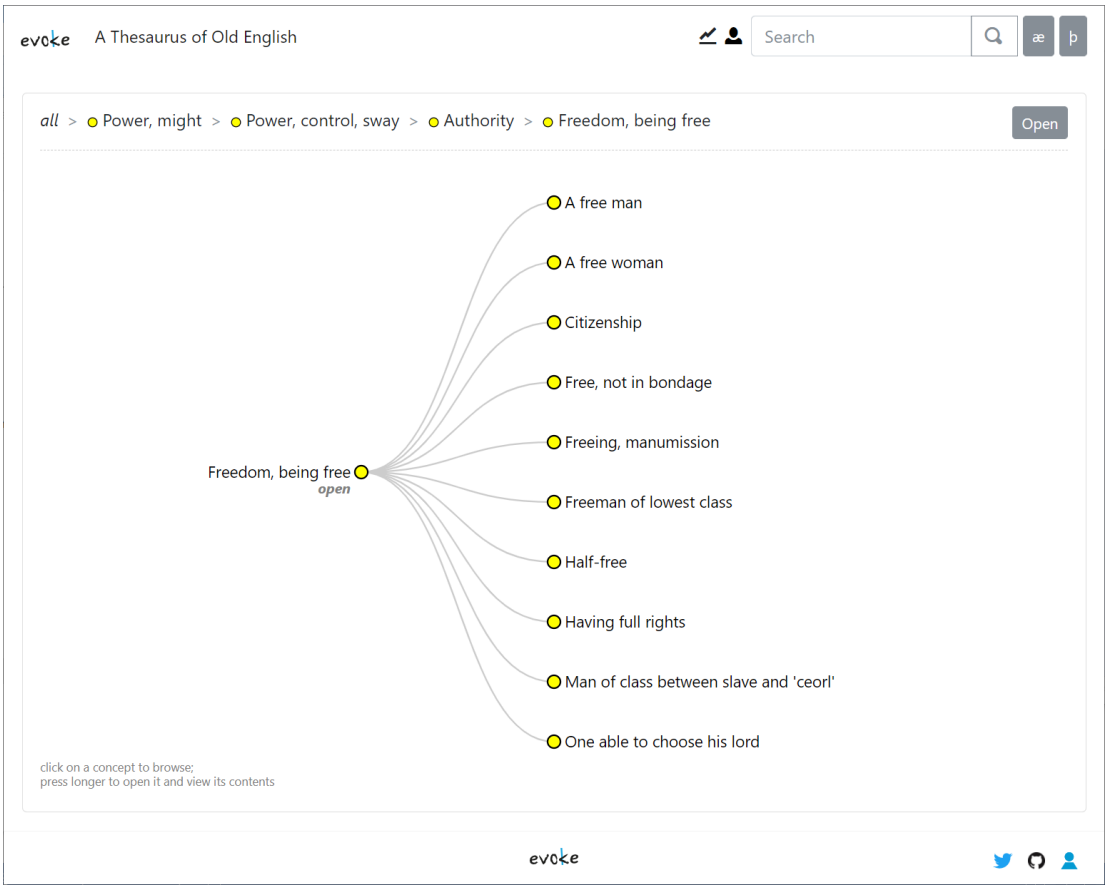


FIGURE 3 Navigating the concept "Freedom, being free" of *A Thesaurus of Old English*

return to a concept higher up in the hierarchy. Semantic concepts of interest can be opened up for viewing (discussed in section 4.2) by clicking on 'open' either next to the breadcrumbs or in the tree-like visualization.

4.1.2 Searching for Content

The search engine of Evoke offers the second manner to approach thesaurus content, accessed through a search box in the navigation bar. This component performs searches for semantic concepts, words, word senses, and any other elements (e.g., labels) based on an entered string. Support for wildcard searches is in place: a question mark (?) allows for any character to occupy that position of the query; an asterisk (*) allows for any number of characters to exist at that position. Thus, searching for *scip** in *A Thesaurus of Old English* will result in a number of compound words that start with the element meaning "ship" (see Figure 4), including the nouns *scipāc* 'oak for shipbuilding' and *scipgyld* 'ship-tax'. Wild card searches enable a whole range of research avenues, including identifying Old English kennings – i.e., poetic compounds such as *sǣwudu* 'ship, lit. seawood' – and for looking into the productivity of affixes (e.g., the suffix *-lēas* '-less' with adjectives).

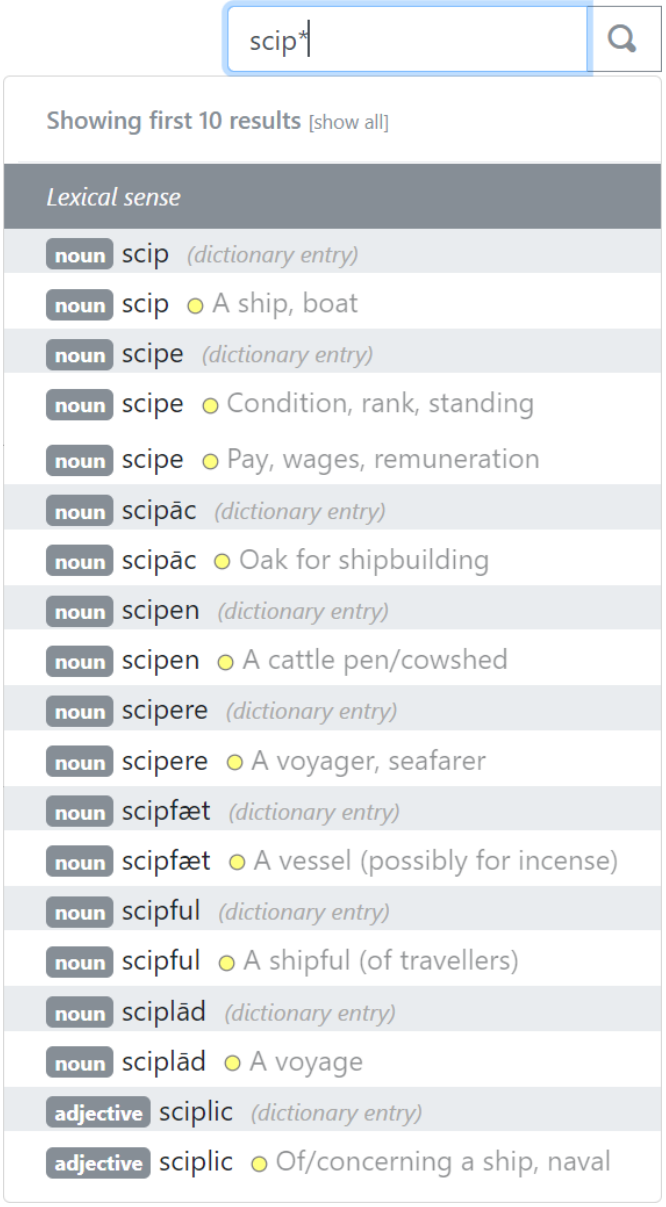


FIGURE 4
Search results for scip* in *A Thesaurus of Old English*

Search results in Evoke indicate the type of each find: semantic concepts are clearly marked and distinguished from, for instance, lexical items or labels. To illustrate, the dark grey bar in Figure 4 acts as heading for the results underneath, all of which are senses. Word senses that belong to the same dictionary entry are grouped together, allowing users to select the word as well as one of its specific senses for viewing. For each lexical item, the lemma is preceded by its part of speech (e.g., "noun") and followed by any labels that are applicable (e.g., "poetry"). As the meanings of word senses are indicated by their location in the semantic hierarchy, word senses in the search results are shown along with the semantic concept that express their meaning (e.g., 'A voyage' for *sciplād*). This presentation enables users to distinguish different senses of the same word from one another.

4.2 *Viewing*

On opening a word, word sense, semantic concept, or any other type of element, for viewing, the earlier discussed navigation components are minimized and, instead, an information pane takes centre stage (cf. R2). The pane contains a number of tabs, each offering a distinctive viewpoint on the element in question. The "info" tab presents basic information on the element viewed: its internationalized resource identifier (IRI), name, type (e.g., concept, word sense, etc.), and other properties available (see Figure 5). For lexical items (i.e., word and word senses) and semantic concepts, the information pane contains a number of additional tabs that incorporate important viewpoints on these Linguistic Linked Data elements specifically.

4.2.1 Viewing Semantic Concepts

An information pane on semantic concepts contains four additional tabs: "list", "statistics", "wordcloud", and "associations". The first of these records the word senses that denote (or lexicalize) the concept that is being viewed. The concept "12.01.01.10 Freedom, being free" in *A Thesaurus of Old English*, for instance,

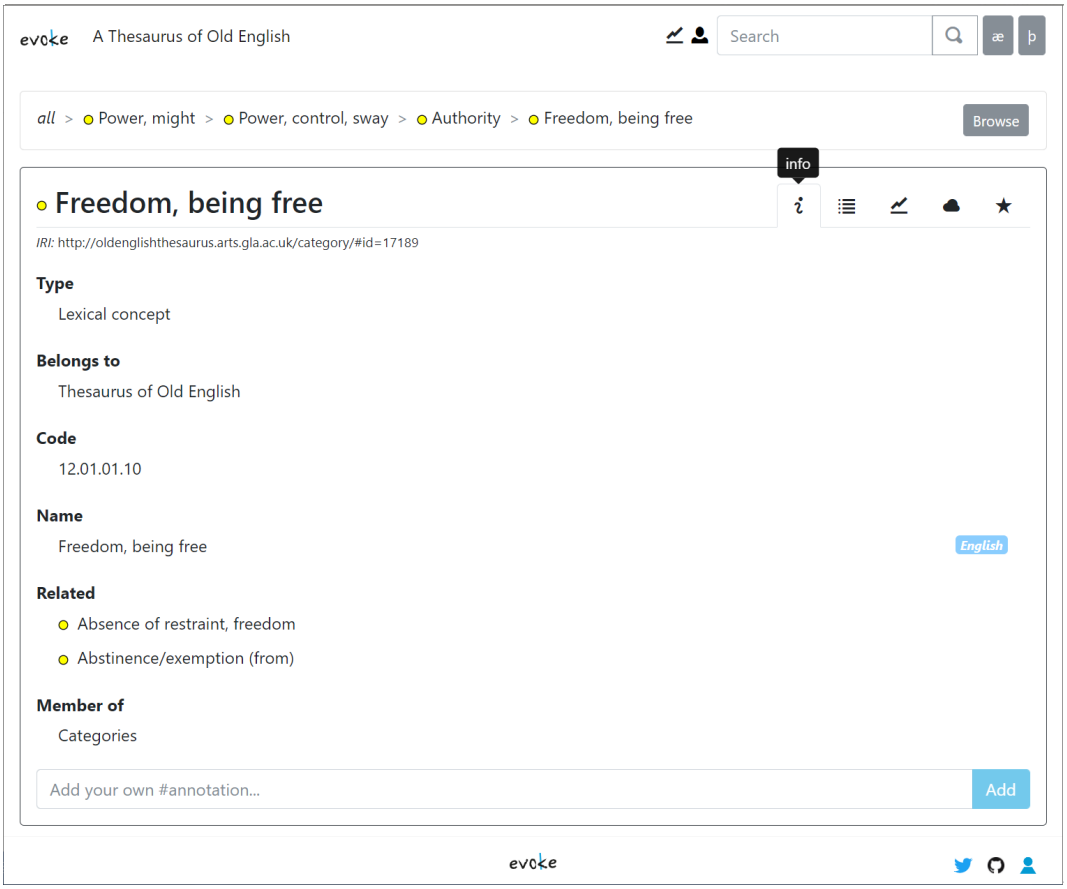


FIGURE 5 Information tab of the concept "Freedom, being free" from *A Thesaurus of Old English*

is denoted by senses attributed to the Old English words *frēols* and *frēot* (see Figure 6). Each word sense listed indicates the part of speech, language (i.e., Old English in *A Thesaurus of Old English*) and any labels that are applicable either to this particular sense or, more generically, the dictionary entry of the word (e.g., "poetry"). Any of these elements can be clicked for viewing instead of the semantic concept currently shown, utilizing the underlying Linked Data mechanisms.

The "statistics" tab provides functionality for analyses and will be covered in section 4.4. The "wordcloud" and "association" tabs present graphical overviews. The former depicts, in a more playful manner, the same words listed on the "list" tab (see Figure 7). The latter conveys semantic concepts that are evoked by further senses of the words found in this semantic domain, indicating possible connotations captured through polysemy (see Figure 8).

197

7

4.2.2 Viewing Lexical Items

Words and word senses have a similar representation to that of semantic concepts. Figure 9 contains an information pane for the Old English word *þēof* in the sense of 'A robber, thief'. The "info" tab, here, includes synonyms, information on the dictionary entry of the word (in a grey box on the right), and an annotation component (at the bottom, discussed in section 4.3.1). The "list" tab for these items indicates the various polysemous senses available for this word and their place in the semantic hierarchy of the thesaurus. The "wordcloud" tab, again, depicts these listed items in a cloud of words, and can be considered a narrower view of the "associations" tab for semantic concepts – here restricted to a single word rather than all those found in a given semantic domain.

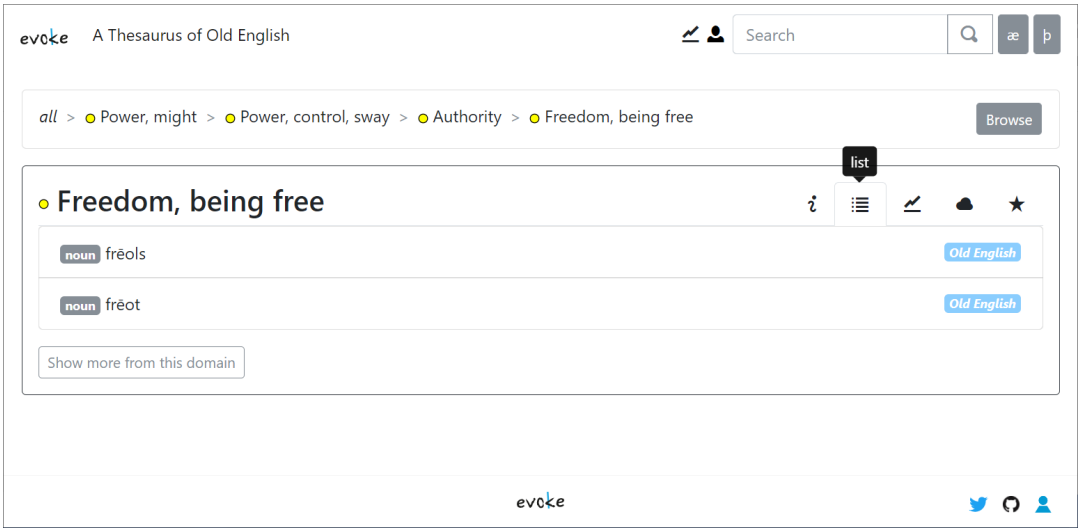


FIGURE 6 List tab of the concept "Freedom, being free" from *A Thesaurus of Old English*

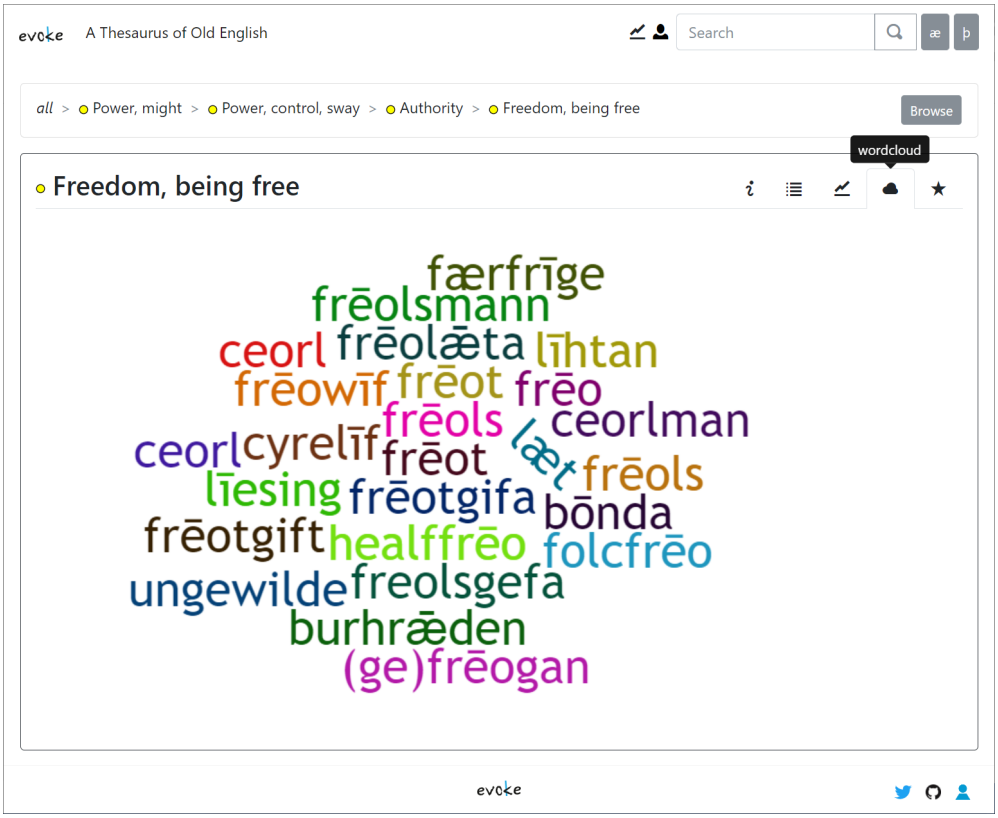


FIGURE 7 Wordcloud tab of the concept "Freedom, being free" from *A Thesaurus of Old English*

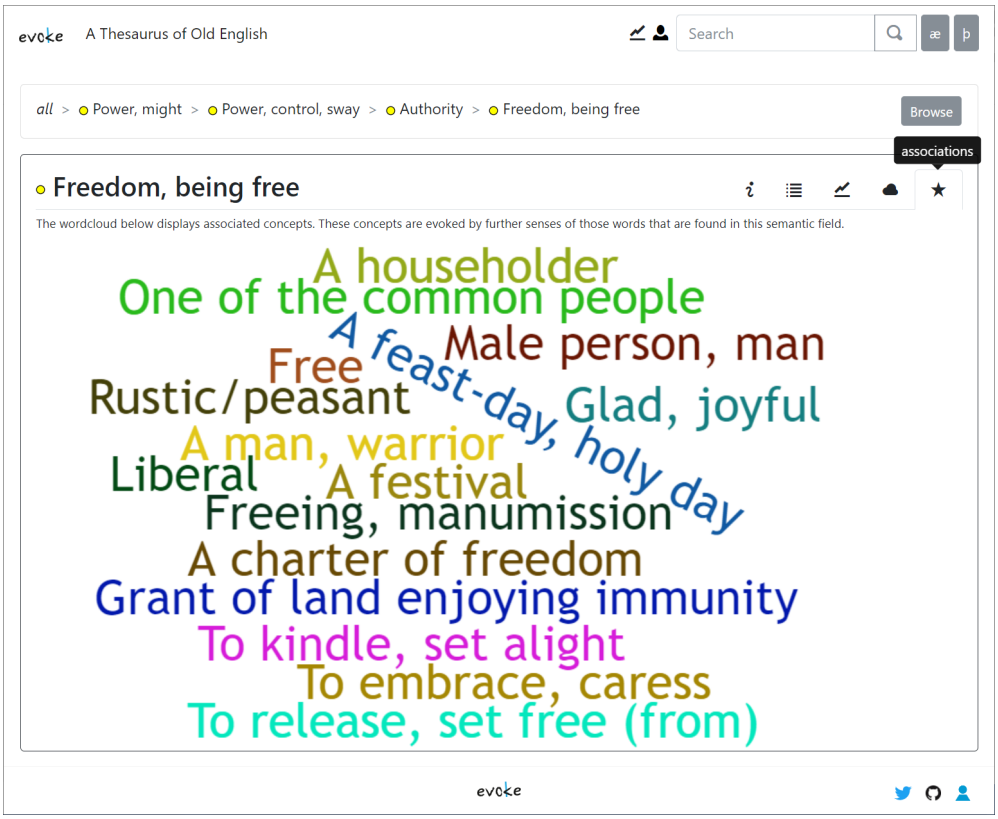


FIGURE 8 Associations tab of the concept "Freedom, being free" from *A Thesaurus of Old English*

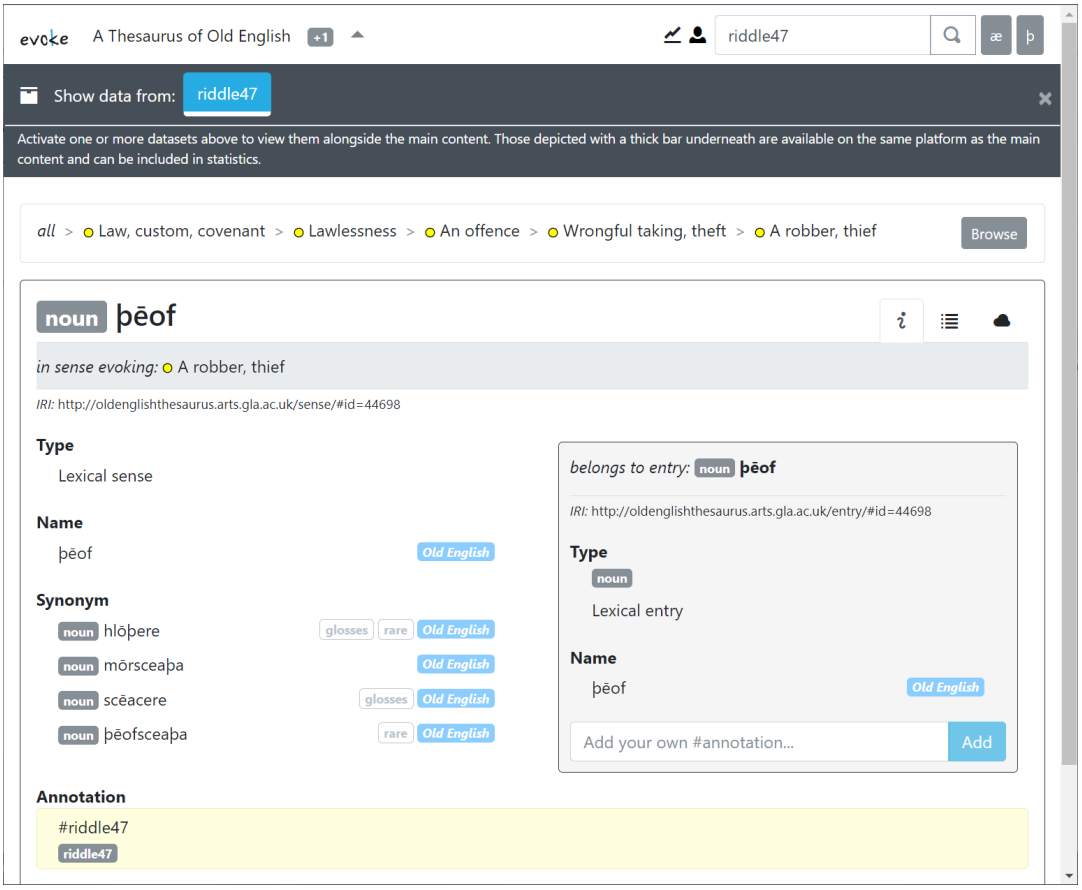


FIGURE 9 A sense of the word *þēof* as described in two datasets (i.e., *A Thesaurus of Old English* and *riddle47*, see top of image)

4.3 *Extending*

One of the main requirements for Evoke has been to allow users to extend thesaurus content, thereby adding salient information (cf. R3). Offering such functionality has been done in various shapes before, but tends to require user accounts, online hosting, moderation of user content, and, possibly, users first obtaining a license for full access to the original dictionary data instead of being able to directly engage with the content at hand (cf. AR3). Evoke takes a different approach, utilizing Linked Data in order to bring together thesaurus content and user content. Users can extend content made available in Evoke in two ways: (1) annotating and (2) linking data.

4.3.1 Annotating

Users of Evoke can annotate content and add their own labels, allowing them to mark words, or their senses, as being noteworthy in a given manner. A given word can, for instance, be labelled *poetic* if its use is mostly restricted to poetry. Further usage features that can be marked though such a system of tags or labels include those based on time (diachronic), place (diatopic),

formality (diaphasic), and frequency (diafrequential) (Hausmann, 1989: 651–652). Marking words in this manner is known in lexicography as diasystematic labelling (Hartmann and James, 1998). Labels are useful beyond recording usage features, however. Indications of occurrences in a specific text or use by a certain author can facilitate systematic analyses of the vocabulary employed within this context. See research done by Thijs Porck and by Amos van Baalen, part of this special issue, which showcase the application of such textual and authorial labels.

The annotation component is presented on information tabs of lexical elements (see Figure 9). Although typically used for annotating words and word senses, the component is not limited to use within that scope and could be used for semantic concepts or indeed any other kind of element. The use of the hashtag symbol (#) in an annotation results in the creation of a corresponding label (e.g., 'riddle47' in Figure 10). Unlike in many existing systems, annotations and labels created in this manner are stored not online in a database but, instead, in the user's Internet browser.[16] Additions and remarks belonging to the user can thus be shown and navigated, but are not disseminated or shared from the outset. In practice, this kind of use of browser storage for annotating can be likened to the practice of scribbling in the margins of one's own printed copy of a dictionary: the original resource (i.e., the thesaurus available for viewing online) is not affected, but only the copy of that resource displayed to the user. In effect, this leads to the creation of a personal copy – one that can be offered without users needing their own user account (cf. AR2).

Users of Evoke are provided full control over their own data. They can download a backup of their annotations and labels in the form of a file, subsequently choose to share these with a select group (e.g., via file sharing or other means), reinsert content in the browser storage, or publish it online for a wider audience to access alongside the thesaurus. As the created backups do not contain the original thesaurus data, but merely the user's own annotations that reference the *identifiers* (or IRIs) of the original thesaurus content, users will not breach any license in place that restrains them from downloading the original dictionary data (cf. AR3). Users simply do not download that data; only their own. Thus, mechanisms offered by Linked Data facilitate in negotiating the interests of both users and publishers.

16 The use of online databases for storing remarks and annotations on lexicographic content can be seen, at the time of writing this article, with resources such as *HTS* (which uses its own database) and *Merriam-Webster* (https://www.merriam-webster.com/, which uses Facebook to store and access comments).

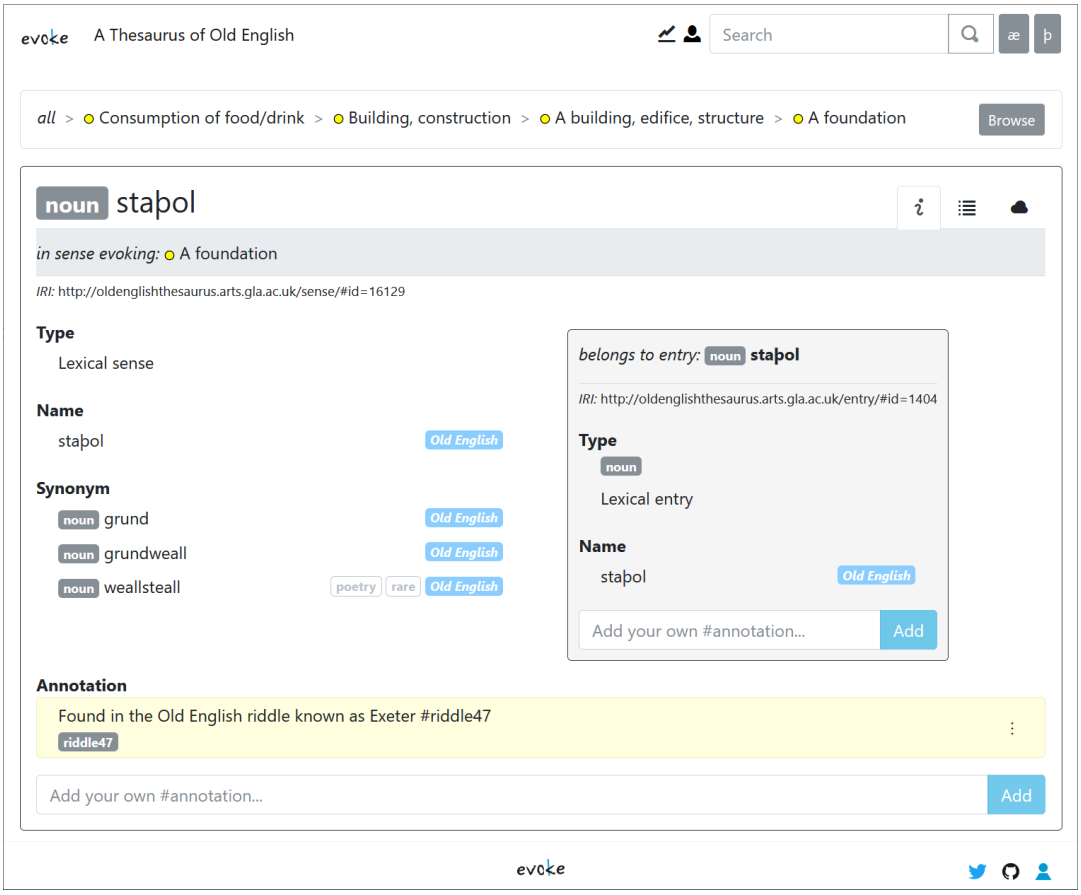


FIGURE 10 An annotation in Evoke (bottom) on the Old English *staþol* in the sense of 'A foundation'

4.3.2 Linking Data

Software other than Evoke can be employed, too, in order to provide additional information on words, word senses, and semantic concepts. The need to facilitate further software to work with and extend existing thesaurus data stems from the fact that some tooling is more suited for a given task (or for a certain audience) than other tooling.[17] Lexicographers may well prefer one piece of software whereas linguists or philologists may turn to another. The use of Linguistic Linked Data, in which content is identified by means of IRIs, offers the possibility to link linguistic content maintained elsewhere by adding references to these internationalized resource identifiers. Indeed, multiple research projects have adopted this approach. Two of the tools that have been used for this purpose, alongside Evoke, are described below. Publications on the research in which these tools have been used are part of this special issue.

17 Separate software has been developed, for instance, for maintaining Linguistic Linked Data content and for creating links between existing datasets in this format. For examples, see section 7.

Further tooling to create and maintain Linguistic Linked Data is actively being developed (e.g., VocBench 3; see Stellato et al., 2020).

In his research project to create textual thesauri of the Old English poems *Beowulf* and *Andreas*, as well as the prose *Old English Martyrology*, Thijs Porck has employed an alignment tool that was specifically created for connecting data from glossaries of Old English texts to the dataset of *A Thesaurus of Old English*, already available in Evoke. The tool reads in a glossary of a text edition (in PDF format), extracts information on the words that occur in that Old English text and presents that information in a table. Each table row offers the user the means to search for, match, and link related content (in this case, words) from *A Thesaurus of Old English* and to annotate these finds with additional remarks or labels (see Figure 11). Indeed, search results in Evoke can be dragged and dropped in other applications, using the mouse cursor, to create a link.[18] The alignment thus created can be exported and/or published in the Linked Data format in order to view the results alongside *A Thesaurus of Old English* in Evoke.

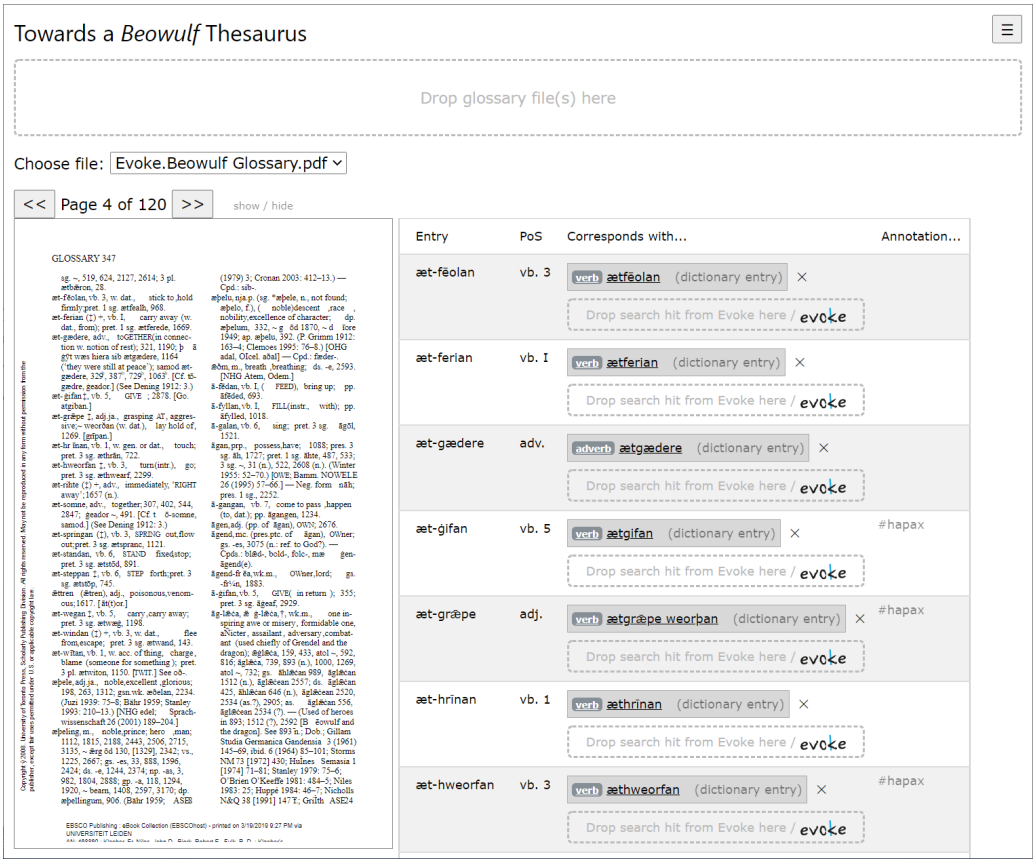


FIGURE 11 Alignment tool used for the Old English text *Beowulf*

18 The drag and drop functionality in Evoke employs the HTML Drag and Drop API (https://html.spec.whatwg.org/multipage/dnd.html#dnd). Thus, the identifiers (or IRIS) of a dragged element from search results in Evoke can be dropped elsewhere.

In research by Rita van de Poel and Sander Stolk, dictionary data on Old Frisian is linked to *A Thesaurus of Old English*. A spreadsheet has been used to store information on Old Frisian words and to categorize their senses according to the semantic hierarchy found in the thesaurus (Figure 12). Each row in the spreadsheet contains information on the sense of an Old Frisian word, which includes the identifier of the word sense (column A), dictionary entry (B), headword (D), language tag (E; here 'ofs' for Old Frisian), but also the identifier of semantic concepts that best express their meaning (F). The latter are either IRIS of semantic concepts found in *A Thesaurus of Old English* or numbers that refer to newly coined ones specific to Old Frisian. In this project, information from the spreadsheet has been transformed to Linguistic Linked Data through the open-source tool OpenRefine along with its RDF plugin.[19]

4.4 *Analysing*

Thesauri are valuable in analysing the vocabulary of a language community. Scholars have made use of these lexicographic works to investigate a range of aspects encoded in the lexicon: cultural elaboration, semantic domains and their cultural connotations, stylistic preferences of authors or in certain texts, and the use and development of metaphors (e.g., Spevack, 1993; Crystal, 2014; Anderson, Bramwell, and Hough, 2016; Porck, 2016: 59–71, 239–294; Diller, 2017).[20] Functionality to perform such analyses is therefore valued greatly for research based on thesaurus content. A key requirement for Evoke in opening up thesauri, then, has been to allow users to perform statistical analyses (cf. R4). The analyses should work on the original thesaurus content but, with the ability of users to extend the thesaurus with their own information, also on

	A	B	C	D	E	F
1	id	entry-id	[link]	word	language	concept-id
2	aftberen-s1	aftberen	link	aftberen	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1120
3	aldaêm-s1	aldaêm	link	aldaêm	ofs	14
4	aldafeder-s1	aldafeder	link	aldafeder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1183
5	aldemôder-s1	aldemôder	link	aldemôder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1184
6	alderlâs-s1	alderlâs	link	alderlâs	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1124
7	alder-s3	alder	link	alder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1183
8	alder-s4	alder	link	alder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1185
9	alder-s5	alder	link	alder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1109
10	âthum-s1	âthum	link	âthum	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1262
11	âthum-s2	âthum	link	âthum	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1262
12	basterdbrôther-s1	basterdbrôther	link	basterdbrôther	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1143
13	basterd-s1	basterd	link	basterd	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1141
14	basterdswester-s1	basterdswester	link	basterdswester	ofs	3
15	bênene burch-s1	bênene burch	link	bênene burch	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1787
16	bêneteburch-s1	bêneteburch	link	bêneteburch	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1787

FIGURE 12 Excel spreadsheet used for dictionary data on Old Frisian

19 OpenRefine, https://openrefine.org/.

20 *HTE* provides a useful overview of publications that utilize its thesaurus, see https://ht.ac.uk/bibliography/.

additional content created by users. This section treats two manners in which users of the Evoke interface can perform analyses: (1) viewing default analyses and (2) building custom queries.[21]

4.4.1 Viewing Default Analyses

Analyses of words in semantic domains are a common use of thesauri. Indeed, in her review of *A Shakespeare Thesaurus*, Christian Kay mentions that a simple count of word senses under a category is sorely lacking in that lexicographic work (Kay, 1996: 73). Such counts indicate the degree of lexicalization, also known as cultural elaboration, of semantic concepts (Wierzbicka, 1997: 10–11). The underlying hypothesis for the importance of these figures is that domains that are important in a culture are heavily encoded in the language of that community, providing its speakers with a multitude of nuances to discuss the subject. This statistic is, for thesaurus content in a digital environment, relatively straightforward to obtain through queries. Even so, many thesauri – both paper and electronic editions – do not include these rudimentary statistics in their presentations.[22]

The web application Evoke presents a small set of basic, default analyses under the statistics tab of semantic concepts. Figure 13 illustrates these analyses for the concept "12.01.01.10 Freedom, being free" of *A Thesaurus of Old English*. Three charts are shown on this tab. The first is a bar chart that indicates the degree of lexicalization. The viewed concept has 2 words specifically in the sense of "freedom, being free" and 28 further words in a sense that is more specific in meaning (i.e., positioned at subordinate concepts in this branch of the semantic hierarchy).

The second chart is a pie chart that indicates the distribution of word senses in this domain – including senses positioned lower down the hierarchy – over parts of speech. The chart in Figure 13 indicates that most words in this domain are nouns (70%) with adjectives and verbs making up the remainder. As Jan Anward points out, words in a part of speech not only share grammatical properties, but tend to belong to the same semantic category, too (Anward, 2006). Nouns, he argues, denote entities, whereas verbs denote events. Linguists who analyse a semantic domain tend to discuss the parts of speech for the lexis involved or concentrate on a single part of speech.[23] Providing statistics on the distribution of parts of speech is therefore warranted for users of Evoke.

21 Users familiar with SPARQL, the standardized querying language for Linked Data, can also formulate queries beyond those available through the Evoke interface in the case of datasets made available through a SPARQL endpoint (see Harris and Seaborne, 2013).

22 To illustrate, the paper and online editions of both *A Thesaurus of Old English* and the *Historical Thesaurus of English* do not display these figures either.

23 E.g., Vea Escarza (2021), who examines Old English verbs of envy.

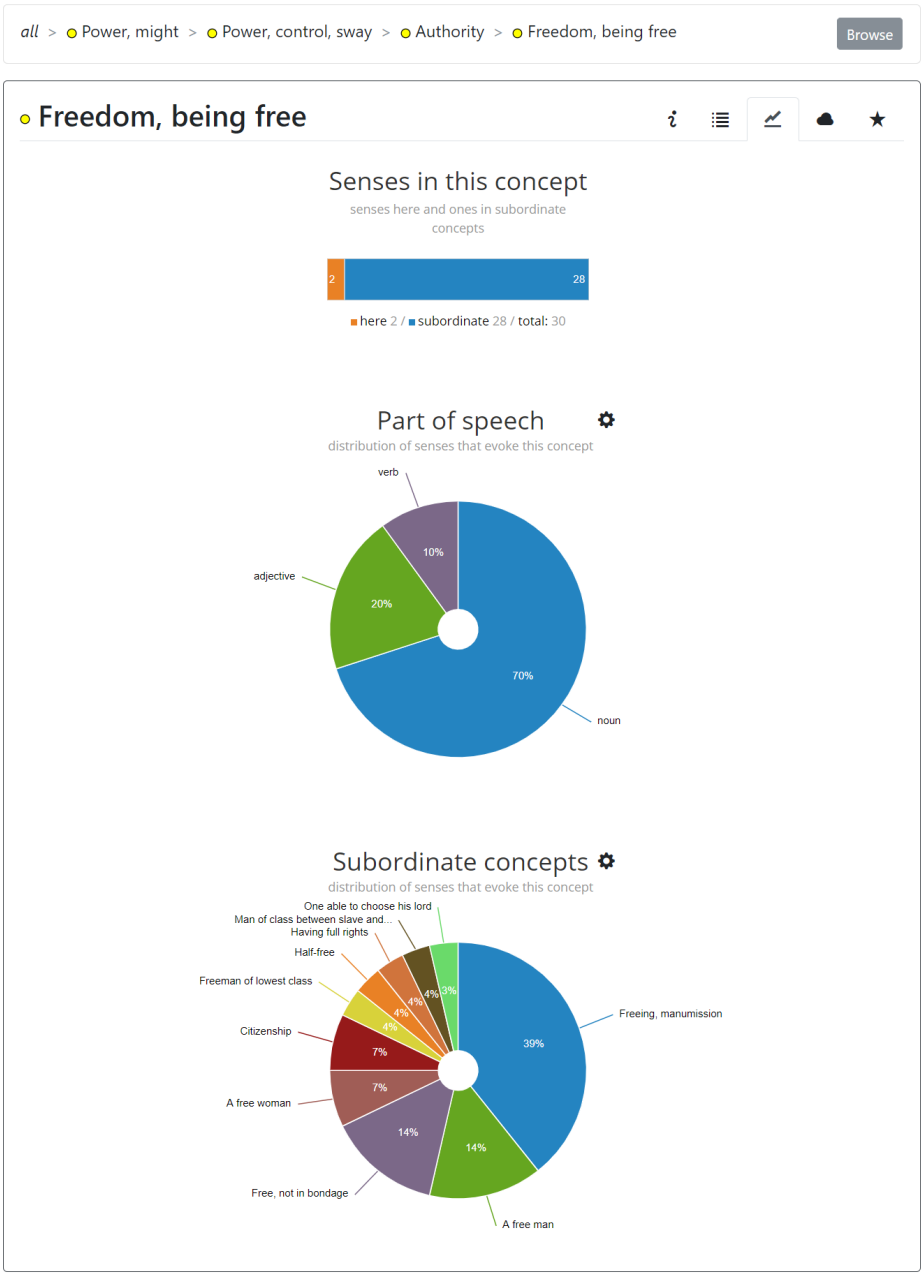


FIGURE 13 Default statistics as shown in Evoke for the semantic concept "Freedom, being free"

The third and last analysis provided by default is the distribution of word senses over the various direct subdomains of the current semantic concept. In other words, this analysis contrasts the various branches from this point in the semantic hierarchy. To illustrate, the chart presented in Figure 13 conveys that there are twice as many words available to express 'Freedom, being free: A free man' than 'Freedom, being free: A free woman'. Hovering over these slices with the mouse cursor presents their absolute counts (in this case, 4 and 2 respectively) instead of their percentages. In contrasting related domains such as these, users are able to analyse their relative degree of lexicalization and therefore supposed dominance or importance in the language community. The larger the semantic

domains analysed, the more valuable it will be for researchers to have these statistics automatically generated as opposed to calculating them manually. In *A Thesaurus of Old English*, for instance, 1,348 word senses evoke the concept "13.02 War", whereas the nine co-ordinate domains for peace encompass a mere 119 word senses. A cog wheel shown next to the pie charts allows users to customize these charts by hiding certain slices. Disposing of "13.02 War" thus, visually, would allow one to foreground "peace" and the nuances between the nine co-ordinate domains that deal with this notion.

4.4.2 Building Custom Queries

On the analysis page of Evoke, accessible by means of the graph icon in the navigation bar, the interface presents users with a way to build custom queries and perform analyses (see Figure 14). Users can indicate which items are of interest by selecting features of words or words in a specific sense (called "lexical entries" and "lexical senses" respectively). Features of these items can be selected afterwards: a label, a part of speech, and, in the case where multiple are present, a language. Although a selection can include only one part of speech and one language, it may contain multiple labels to combine features of interest (e.g., "poetry" and "rare" for rarely occurring items found in poetic texts). Possible values for a given feature are presented through a dropdown list. Each option in the list is presented as text.[24]

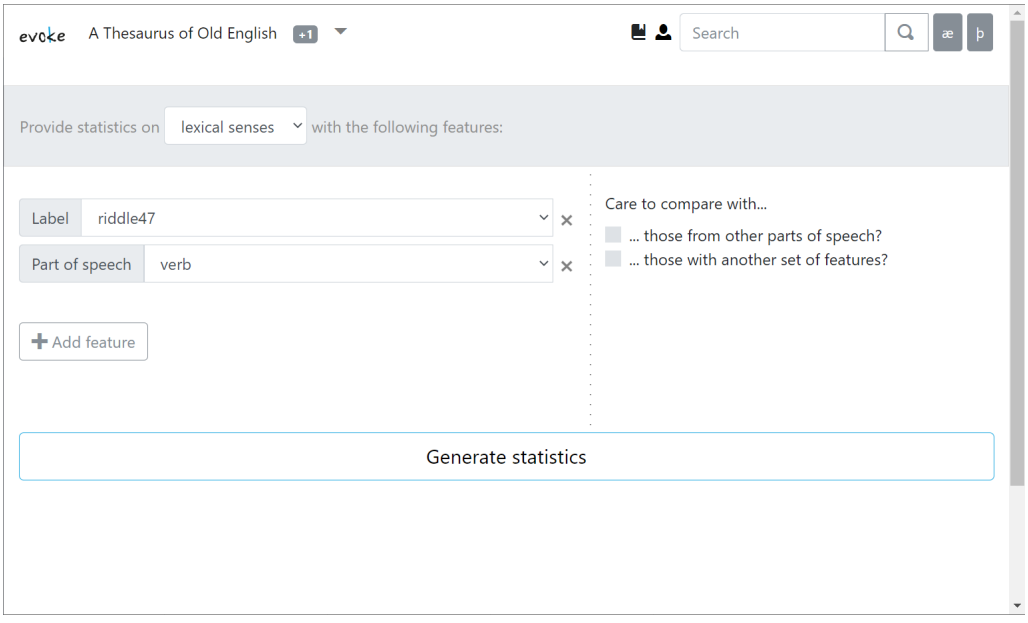


FIGURE 14 The form that allows users to select their features of interest in the statistics section

24 A tooltip – shown upon hovering over an option with the mouse – indicates the IRI or language code that identifies that feature in Linked Data.

Once one or more features have been selected for analysis, users can opt to compare their selection with items from other languages (if any are available in the data analysed), with other parts of speech, or with another set of features they are interested in. The default choice in analyses is to contrast items of interest with all other items of that kind. Thus, a selection of lexical entries with the labels "poetry" and "rare" would be contrasted with all known lexical entries, regardless of their features. Once a user has formulated their query, statistics can be generated and presented. Details on the selection made are stored as part of the URL (as query string arguments), allowing users to share these results by copying the current Web address from their Internet browser.

Analysis results are shown as charts. Each chart presents a distinctive view on the selection made (see Table 1). Together, the charts represent fundamental analyses that take into account the organization of words and word senses in the overarching semantic hierarchy of the thesaurus, thus offering what we term an onomasiological profile of the selection made.

Additionally, the visualisation of most charts (i.e., all but the item count) can be customized to suit research needs. Data series shown can, for instance, be enabled or disabled by clicking on their name in the legend. The x-axis of

TABLE 1 Charts available for an onomasiological profile based on a user-specified selection

Chart	Description
Item count	This chart shows the number of items in the selection.
Degree of ambiguity	This chart shows the number of lexical entries based on their sense count, also known as the degree of polysemy. (If the selection criteria are based on lexical senses, this chart plots the entries to which they are attributed.)
Degree of synonymy	This chart shows the number of lexical senses based on their synonym count. (If the selection criteria are based on lexical entries, this chart plots the senses attributed to them.)
Distribution: categories	This chart shows the distribution of lexical senses over categories, see Figure 15. (If the selection criteria are based on lexical entries, this chart counts the senses attributed to these entries.)
Distribution: tree depth	This chart shows the distribution of lexical senses over the various levels of the taxonomy, see Figure 16. (If the selection criteria are based on lexical entries, this chart counts the senses attributed to these entries.)

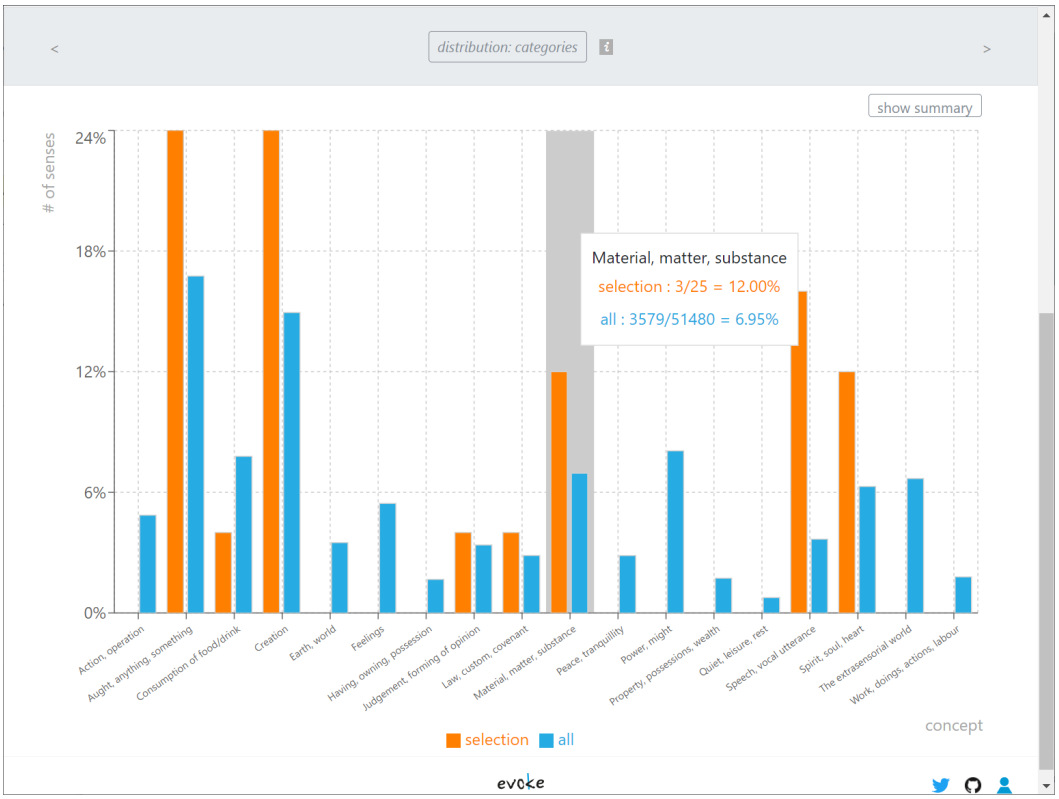


FIGURE 15 The distribution over semantic concepts of word senses labelled 'riddle47' (orange) versus all senses (blue)

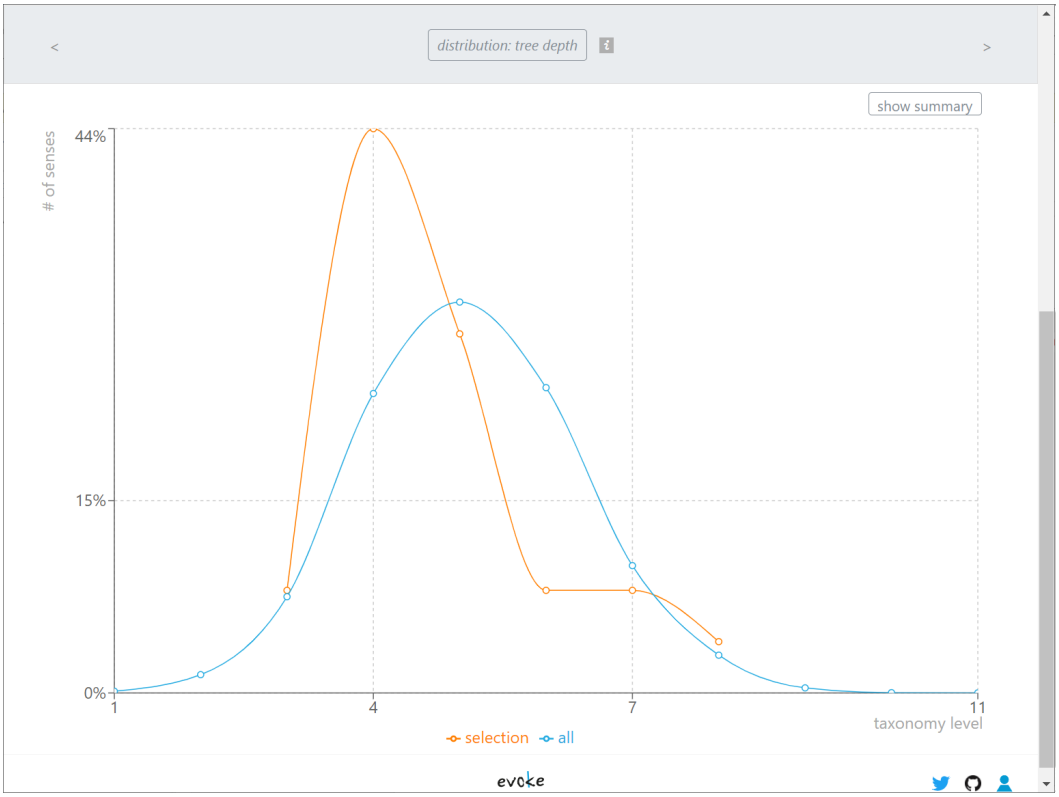


FIGURE 16 The distribution over taxonomy depth of word senses labelled 'riddle47' (orange) versus all senses (blue)

line charts can be zoomed in through clicking and dragging the mouse over the desired range. The y-axis of charts can be altered, by clicking on one of its value labels, to present values as absolute numbers or as percentages indicating the values relative to the cumulative value of the data series. Moreover, charts can include only those items that are (or have) a lexical sense located *in a given branch* of the thesaurus taxonomy. Selecting one of the categories in the chart covering the distribution of items over categories will allow the user to dive into that branch of the taxonomy. The resulting cropped selection – applied to all charts – is marked through a bar above the charts. Finally, a table containing summary statistics (average, median, and domain and range of the chart) is available through the button 'show summary', located in the top right corner.

5 *A Thesaurus of Old English* as Linguistic Linked Data

The first lexicographic resource made available in Evoke as Linguistic Linked Data (LLD) is *A Thesaurus of Old English* (henceforth *TOE*). This thesaurus captures the lexis of Old English, the early medieval variant of English spoken between roughly 500 and 1100. Upon its first publication in 1995, this resource has been met with high praise, having been called "the most important contribution to Old English studies for years," since the "comprehensive analysis" that it establishes allows scholars to "investigate what distinctions Anglo-Saxons felt important enough to make in the lexicon" (Görlach, 1998: 398).

The LLD version of *TOE* (henceforth *TOE*-LLD) has been available in Evoke since 2018, based on an extract of the *TOE* database on May 26th, 2017 that was kindly provided by the University of Glasgow. The majority of the elements within the database were straightforward to translate into Linguistic Linked Data. The Modern English "categories" in *TOE*, for instance, correspond with "lexical concepts" in LLD; "lexemes", which have been allocated to *TOE* categories in the extract of the database, correspond with a LLD "lexical sense" (i.e., a word in a specific sense) that "lexicalizes" a concept.[25] For details, see Stolk (2019).[26] One aspect in transforming *TOE* to *TOE*-LLD has been altered since that publication and warrants further explanation.

The original database of *TOE* does not group different senses of a single word under a single entry, a practice typical in dictionaries. Even so, research can benefit from such lexical entries. A case in point is *A Beowulf Thesaurus*,

25 See terms LexicalConcept, LexicalSense, and isLexicalizedSenseOf in OntoLex (Cimiano, McCrae, and Buitelaar, 2016).

26 The source code developed for performing the translation is available at https://github.com/ssstolk/toe-lld/.

which marks words used in the Old English text *Beowulf* (see the contribution by Thijs Porck in this special issue). Marking word senses rather than words, instead, would have constituted an interpretive act which can be likened to translation of the text. Selected senses may not coincide with the author's original intentions, and would discard some of the ambiguity that the poetic text holds. This interpretive act is one that was avoided in the creation of *A Beowulf Thesaurus*, but demanded lexical entries to be present in *TOE*-LLD. Thus, words rather than separate word senses could be marked as occurring in *Beowulf*. Moreover, distribution flags in *TOE* "relate only to word forms, not to meanings".[27] As a consequence, these flags convey information beyond the level of a lexical sense and can be considered to belong to lexical entries instead, albeit ones implicit in the original database.

In order to enrich *TOE*-LLD with entries, senses were grouped automatically if they had the following in common: their headwords, their parts of speech, and the distribution flags attributed to them (e.g., "p" for items found only in poetry, "g" for those found only in glosses; see contribution by Jane Roberts in this special issue). Based on these features, "lexical entries" were created in *TOE*-LLD. As contributors to this special issue point out, the automated detection of entries is by no means perfect: *TOE* occasionally records separate word forms for a single lemma (e.g., *gēardagas* and *gēardagum*), homonyms may be grouped together (e.g., *dung* meaning "dung, manure" and "dungeon, prison"), and, although rare, distribution flags may not be attributed consistently for different senses of the same lemma (see the contribution by Porck to this special issue). Addressing these issues will require manual verification in order to merge or separate entries that have been created.[28]

Evoke offers some functionality for interacting with *TOE* beyond that provided by the existing website of the thesaurus hosted by the University of Glasgow. The original website lacks the means to extend the thesaurus, due partly to licensing concerns. In contrast, Evoke allows for the extension of the thesaurus content by annotating and creating links, which is facilitated by the LLD version of *TOE*. Not only categories have an IRI in *TOE*-LLD, as is the case on the original website, but every element is identified by one: from categories to lexical senses and labels. Moreover, Evoke offers the means to perform automated analyses on the information available, based on the semantic hierarchy of the thesaurus, resulting in onomasiological profiles.

27 https://oldenglishthesaurus.arts.gla.ac.uk/flags/.

28 Technically, adding the distinction between lexical senses and entries could be added to the original database of *TOE*. Moreover, recording entries in this database would remove redundancy and keep distribution flags consistent (see Stolk, 2019).

6 Evoke and Old English Studies

To assess the usefulness of Evoke both in research and for education, a research project was formulated with the title 'Exploring Early Medieval English Eloquence' (EEMEE). This project has brought together seventeen scholars from universities and lexicographic institutions from across Europe to explore the contents of *TOE* using Evoke, from a variety of disciplinary perspective, ranging from linguistics to literary-criticism, history, lexicography, and philology. In their explorations, the researchers (and in educational settings, their students) set about viewing other material next to that of *TOE*, some by linking an already existing source and others by using the annotation system in Evoke. These researchers have been able to access and extend the thesaurus through the use of Evoke, performing advanced analyses, whilst abiding by the license of the lexicographic resource, which restricts them from extracting or downloading substantial portions of the dataset. The results of some of their research projects are included in this special issue.

Novel research done within EEMEE includes analyses of lexis used in specific texts (e.g., *Beowulf*, see Thijs Porck in this special issue) or used by specific authors (e.g., Ælfric, see Amos van Baalen in this special issue). Their analyses have generated new insights into semantic choices made by Anglo-Saxon authors in individual texts or across an entire oeuvre. The onomasiological profiles thus established may act as semantic fingerprints that can be used in comparative analyses. Other research has focused on metaphors associated with anger and their development through the history of the language (see Khan et al. in this special issue). Lastly, a number of researchers have worked on linking Old Frisian and Old Dutch lexis on kinship with the semantic hierarchy of the thesaurus, which allows us to contrast how many nuances these language communities respectively had, next to those of Old English, in expressing such concepts (see contributions by Rita van de Poel and Sander Stolk; and by Katrien Depuydt and Jesse de Does). All the aforementioned researchers have linked up additional information to the original thesaurus content.

As regards the usefulness of thesauri in education, in the past two years students have used Evoke and *TOE* to explore aspects of Old English language and culture. At the University of Groningen, Evoke has been used as part of an introductory course to Old English (see the contribution by Kees Dekker in this special issue); students at Leiden University participate in a 2-hour workshop that familiarizes them with digital tools and resources for studying Old English: *TOE* and Evoke, alongside the *DOE* and the *DOEC*.[29] The learn-

29 Workshop materials, created by Thijs Porck and Sander Stolk, are available on request.

ing exercises created at both universities are to be incorporated in the Evoke website. By courtesy of Prof. Carole Hough (University of Glasgow), these exercises will include units from the module *Learning with the Online Thesaurus of Old English* (Hough and Kay, 2017).

In short, the dataset of *TOE* has been put to new and innovative use within the field of Old English studies through Evoke. This has allowed the examinations presented in this special issue to provide valuable insights into Old English language and culture. In addition, some of the projects have also led to refinements of the original thesaurus content itself, which were incorporated by the editors of *TOE*, including the insertion of new words and corrections of distribution flags. An unfortunate side-effect of these very recent, minor improvements of *TOE* is that, as of writing this article, the lexicographic resource offered on the website by the University of Glasgow is slightly more up-to-date with the current state of Old English lexicography than the *TOE*-LLD version available in Evoke. Creating a way of automatically synchronizing *TOE*-LLD with the original *TOE* database, in order to avoid discrepancies however minor, is therefore desirable and prioritized on the roadmap for future work surrounding Evoke.

7 Digital Humanities and Onomasiological Research

Digital approaches for investigating various facets surrounding language, including the structure of vocabularies and diachronic perspectives on language development, are by no means new. Such matters have often been explored, in various branches of computational linguistics, with digital corpora and analytical tools (Sula and Hill, 2019). Onomasiological approaches to language, too, have a long history. Early works in which words and phrases have been arranged thematically, rather than alphabetically, date back as far as Antiquity and possibly further still (Hüllen, 1999: 44). Knowledge catalogued by onomasiological works have had numerous uses: interpretation of texts in foreign languages, selection of words or phrases more suitable in textual composition, and studies of entire semantic fields (Hüllen, 1999).[30] Since the last century, research programmes have sought to further harness the potential of digital thesauri, whether digitized or born-digital. Automated uses, taking advantage of their digital form, include natural language processing and

30 Examples of such semantic field studies are, for instance, listed at https://ht.ac.uk/bibliography/.

automated translations.[31] This section positions Evoke within a number of important developments within the field of Digital Humanities that deals with the study of lexicography and onomasiology.

First and foremost, the efforts surrounding Linguistic Linked Data and software developed for exploring resources expressed in this format constitute an important context for Evoke. These efforts, as Declerck et al. (2020) have observed, play "an increasing role in eLexicography" (5664). The English WordNet, for instance, has recently been ported to this model (McCrae et al., 2020). Moreover, several recent initiatives aim at building and maintaining Linguistic Linked Data resources, including the H2020 projects ELEXIS (2018–22), Prêt-à-LLOD (2019–22) and the COST Action NexusLinguarum (2019–23).[32] Tooling in these initiatives that work with Linguistic Linked Data focus on creation, discovery, transformation, and linking (Declerck et al., 2020). Examples of such tools include LingHub, which offers discovery of language resources by searching through their metadata (McCrae et al., 2015), and NAISC, used for aligning two RDF datasets.[33] Unfortunately, most applications currently available for working with Linguistic Linked Data "come with a considerable entry barrier and they address the advanced user of RDF technologies rather than a typical linguist" (Chiarcos et al., 2020). Evoke is amongst the first range of applications that aims to provide a user-friendly interface for such resources and to open them up to a wider audience. Other notable applications that provide user interfaces for Linguistic Linked Data resources are VocBench 3 and LexO (Stellato et al., 2020). Both of these web-based platforms allow users to edit and view Linguistic Linked Data in a user-friendly manner. However, unlike Evoke, they lack functionality to perform onomasiological analyses: their main aim is to manage and publish content collaboratively.

Evoke can also be compared to software capable of working with Linked Data in general or, more specifically, with indexing thesauri expressed in that format, which is more prevalent than that for Linguistic Linked Data. WebProtégé and TopBraid are examples of tools that allow users to edit the graph-like structure of Linked Data through a user interface (Tudorache et al., 2013).[34] Thesauri adhering to the SKOS vocabulary in Linked Data, which will henceforth be referred to as indexing thesauri, tend to resemble the semantic hierarchy of a lexicographic thesaurus: they identify concepts, possibly

213

7

31 See, for instance, the use of machine-learning based on BERT in Kohli (2021).

32 ELEXIS: https://cordis.europa.eu/project/id/731015, 2018–2022; Prêt-à-LLOD: https://cordis.europa.eu/project/id/825182, 2019–2022; NexusLingarum: https://www.cost.eu/actions/CA18209, 2019–2023.

33 https://github.com/insight-centre/naisc.

34 Topbraid, https://www.topquadrant.com/products/topbraid-enterprise-data-governance/.

arranged in a hierarchy, that are used to index material (e.g., images, documents). PoolParty and Skosmos are instances of web-based tools that allow for editing and documenting indexing thesauri, respectively (Schandl et al., 2010; Suominen et al., 2015). Thesauri represented as Linguistic Linked Data are, not coincidentally, also based on SKOS for their semantic hierarchy (indeed, a 'lexical concept' in OntoLex is a specialization of a SKOS 'concept') but add additional terminology to capture lexical entries and senses, which are effectively indexed through their evoking and lexicalising of concepts (Stolk, 2019).

Evoke also shares some characteristics with software that offers functionality for extending resources on the Web through Linked Data mechanisms. Such functionality, although not applied specifically to Linguistic Linked Data, is pivotal in notable recent work such as the tool hypothes.is,[35] used specifically for annotating webpages, and the ecosystem SOLID, which relies on personal RDF data hubs.[36] Both works use online databases to store information of users, requiring them to login to their account before they can add data. In contrast, Evoke demands no login as user data is stored locally, i.e., in memory of the internet browser, instead. Evoke grants users complete control over their own data and annotations (backup, share, publish), does not demand for that data to be stored online in a centralized manner, and requires no account details before interacting with a resource and extending it. This approach both avoids public comments cluttering webpages of annotated resources and encourages users to engage in open science.

Lastly, a number of recent research programmes have increased efforts that expand the use of thesauri to other domains. The onomasiological lens that *HTE* provides, for instance, has been utilized for mapping metaphors throughout the history of the English language (Anderson, Bramwell, and Hough, 2016) and for semantically annotating entire textual corpora for topical analyses (Piao, 2017).[37] Similarly, the work on Evoke has sought to contribute novel methods to Digital Humanities research for engaging with thesauri. By offering statistical analyses utilizing the semantic hierarchy of these lexicographic resources, and by allowing researchers to link additional information to thesaurus content, Evoke grants new, meaningful insights into a language and the use of its vocabulary in cultural expressions (e.g., individual texts or entire oeuvres). The functionality available offers results that, as a number of researchers point

35 http://hypothes.is/.

36 https://solidproject.org/.

37 The corpora currently tagged are the verbatim reports of speeches in the houses of the British Parliament (see https://hansard.hud.ac.uk) and a substantial portion of Early English Books Online (found tagged on https://www.english-corpora.org).

out in this special issue, provides additional knowledge, but may also raise new questions that warrant a closer inspection of the social context (e.g., textual, historical, socio-economic). Evoke, therefore, is firmly rooted in Digital Humanities, and provides the means to explore Humanities-based questions through digital tools that complement, but not supplant, knowledge and expertise of scholars.

215

7

8 Conclusion

Evoke is one of the first applications that provides a user-friendly interface for working with Linguistic Linked Data resources, opening up *TOE* to users interested in engaging with its lexicon from an onomasiological perspective. The design of this web application ensures users can view, navigate, extend, and analyse thesaurus content together with relevant, linked datasets. In doing so, they are able to explore the vocabulary of a language community and its relation to their culture. The usefulness of attaching salient information to a lexicographic resource such as *TOE* is demonstrated by the research presented in this special issue journal. Each contribution showcases how Evoke allows for novel engagement with the Old English lexicon in exploring and extending *TOE*.

Future work on Evoke will have two main aims. The first is to improve its functionality for research and education using *TOE*. Future functionality most frequently requested by researchers is the ability to filter content when navigating and viewing (e.g., based on user-defined labels). This feature would effectively allow users to create specialized subthesauri. Another request is the means to link attestations in corpora to thesauri content, which would permit incorporating frequency analyses in onomasiological profiles. The second main aim for future work on Evoke is to enhance its uptake for research and use beyond Old English studies. Support for additional linguistic resources (including non-English, historical language, multilingual, and sign language datasets), and possibly also indexing thesauri, is considered key for such developments. Moreover, users should be provided with the means to develop their own thesauri instead of working with ones already published and transformed to Linguistic Linked Data. These future developments are planned through collaborative efforts with researchers and language experts from multiple institutes, across national boundaries.

If the initial work on Evoke is any indication, ventures into Digital Humanities depend on collaboration and creativity of researchers as much as on the digital tools that facilitate novel research queries (or, perhaps, even more so). Since

onomasiological investigations rely on the meanings people attribute to symbols (i.e., words and phrases), this human aspect will always remain important in the computer-assisted exploring, extending, and analysing of thesauri.

Acknowledgement

This work was supported by the LUCAS Extra Resources Open Call-II Grant 2020, awarded by the Leiden University Centre for the Arts in Society, and the LUCDH Small Grant 2018, awarded by the Leiden University Centre for Digital Humanities. The author would like to thank the researchers and students who – through workshops, courses, and research projects – have used Evoke and provided valuable input for its development.

References

Adamska-Sałaciak, A. Review of *Historical Thesaurus of the Oxford English Dictionary*, *International Journal of Lexicography* 23 (2) (2010), 227–233.

Albertoni, R., D. Browning, S. Cox, A. Gonzalez Beltran, A. Perego, and P. Winstanley, eds. "Data Catalog Vocabulary (DCAT) – Version 2: W3C Recommendation 04 February 2020." *W3C* (2020), https://www.w3.org/TR/vocab-dcat-2/.

Anderson, W., E. Bramwell, and C. Hough, eds. *Mapping English Metaphor through Time* (Oxford: OUP, 2016).

Anward, J. "Word Classes/Parts of Speech: Overview." In *Encyclopedia of Language and Linguistics*, 2nd edn, ed. K. Brown (Amsterdam: Elsevier, 2006), 628–632.

Bellandi, A., and E. Giovannetti. "Involving Lexicographers in the LLOD Cloud with LexO, an Easy-to-Use Editor of Lemon Lexical Resources." In *Proceedings of the 7th Workshop on Linked Data in Linguistics* (*LDL-2020*), eds. M. Ionov, J. P. McCrae, C. Chiarcos, T. Declerck, J. Bosque-Gil, and J. Gracia (Marseille: European Language Resources Association, 2020), 70–74, https://aclanthology.org/2020.ldl-1.10.pdf.

Bremmer Jr, R. H. "Treasure Digging in the Old English Lexicon." *NOWELE* 40 (2002), 109–114.

Brewer, C. Review of *Historical Thesaurus of the Oxford English Dictionary*, *The Review of English Studies* 61 (252) (2010), 801–805.

Busse, B. "A Celebration of Words and Ideas: The Stylistic Potential of the *Historical Thesaurus of the Oxford English Dictionary*." *Language and Literature* 21 (1) (2012), 84–92.

Chiarcos, C., B. Klimek, C. Fäth, T. Declerck, and J. P. McCrae. "On the Linguistic Linked Open Data Infrastructure." In *Proceedings of the 1st International Workshop on*

Language Technology Platforms, eds. G. Rehm, K. Bontcheva, K. Choukri, J. Hajič, S. Piperidis, and A. Vasiljevs (Marseille: European Language Resources Association, 2020), 8–15, https://www.aclweb.org/anthology/2020.iwltp-1.2.

Chiarcos, C., et al. "Modelling Frequency and Attestations for OntoLex-Lemon." In *Proceedings of Globalex 2020 Workshop on Linked Lexicography*, eds. I. Kernerman, S. Krek, J. P. McCrae, J. Gracia, S. Ahmadi, and B. Kabashi (Marseille: European Language Resources Association, May 2020), 1–9, https://www.aclweb.org/anthology/2020.globalex-1.1.pdf.

Chiarcos, C., et al. "Towards Open Data for Linguistics: Lexical Linked Data." In *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, eds. A. Oltramari, P. Vossen, L. Qin, and E. Hovy (Heidelberg and Berlin: Springer, 2013), 7–25.

Cimiano, P., C. Chiarcos, J. P. McCrae, and J. Gracia. *Linguistic Linked Data: Representation, Generation and Applications* (s.l.: Springer, 2020).

Cimiano, P., J. P. McCrae, and P. Buitelaar, eds. "Lexicon Model for Ontologies: Community Report, 10 May 2016: Final Community Group Report 10 May 2016." *W3C* (2016), http://www.w3.org/2016/05/ontolex/.

Crystal, D. *Words in Time and Place: Exploring Language through the* Historical Thesaurus of the Oxford English Dictionary (Oxford: OUP, 2014).

Cyganiak, R., D. Wood, and M. Lanthaler, eds. "RDF 1.1 Concepts and Abstract Syntax: W3C Recommendation 25 February 2014." *W3C* (2014), https://www.w3.org/TR/rdf11-concepts/.

Dance, R. Review of *A Thesaurus of Old English*, *Medium Ævum* 66 (2) (1997), 312–313.

Declerck, T., et al. "Recent Developments for the Linguistic Linked Open Data Infrastructure." In *Proceedings of the 12th Language Resources and Evaluation Conference*, eds. N. Calzolari et al. (Marseille: European Language Resources Association, 2020), 5660–5667, https://www.aclweb.org/anthology/2020.lrec-1.695.

Diller, H.-J. "Measuring the Growth of Semantic Fields: The Case of the English Emotion Lexicon." In *Cognition in Language: Volume in Honour of Professor Elżbieta Tabakowska*, eds. W. Chłopicki, A. Pawelec, and A. Pokojska (Kraków: Tertium, 2017), 574–596.

DOE = Cameron, A., A. Crandell Amos, A. diPaolo Healey, et al., eds. *Dictionary of Old English: A to I online* (Toronto: Dictionary of Old English Project, 2018). http://tapor.library.utoronto.ca/doe/.

DOEC = DiPaolo Healey, A., with J. Price Wilkin and X. Xiang, ed. *Dictionary of Old English Web Corpus* (Toronto: Dictionary of Old English Project, 2009), https://tapor.library.utoronto.ca/doecorpus/.

Evoke = Stolk, S. Evoke (Web application, 2018), http://evoke.ullet.net/.

Geïntegreerde Taalbank (Leiden: Instituut voor de Nederlandse Taal, 2018), https://gtb.ivdnt.org.

Görlach, M. Review of *A Thesaurus of Old English*, *Anglia* 116 (3) (1998), 398–401.

Harris, S., and A. Seaborne. "SPARQL 1.1 Query Language: W3C Recommendation 21 March 2013." *W3C* (2013), https://www.w3.org/TR/sparql11-query/.

Hartmann, R. R. K. "Thesauruses." In *Encyclopedia of Language and Linguistics*, 2nd edn, ed. K. Brown (s.l.: Elsevier, 2006), 668–676.

Hartmann, R. R. K., and G. James. "Diasystematic Labelling." In *Dictionary of Lexicography* (London: Routledge, 1998), 40.

Hausmann, F. J. "Die Markierung im Allgemeinen Einsprachigen Wörterbuch: Eine Übersicht." In *Wörterbücher: Ein internationales Handbuch zur Lexikographie*, eds. F. J. Hausmann, O. Reichmann, H. E. Wiegand, and L. Zgusta (Berlin: Walter de Gruyter, 1989), vol. 1, 649–657.

Hickson, I., ed. "Web Storage (Second Edition): W3C Recommendation 19 April 2016." W3C (2016), https://www.w3.org/TR/webstorage/.

Hough, C., and C. Kay, eds. *Learning with the Online Thesaurus of Old English* (*TOE*) (Glasgow: University of Glasgow, 2017), http://oldenglishteaching.arts.gla.ac.uk/.

HTE = Kay, C., M. Alexander, F. Dallachy, J. Roberts, M. Samuels, and I. Wotherspoon, eds. *The Historical Thesaurus of English*, 2nd edn, version 5.0 (Glasgow: University of Glasgow, 2021), https://ht.ac.uk/.

HTOED = Kay, C., J. Roberts, M. Samuels, and I. Wotherspoon, eds. *Historical Thesaurus of the Oxford English Dictionary* (Oxford: OUP, 2009).

HTS = Rennie, S., ed. *Historical Thesaurus of Scots* (Glasgow: University of Glasgow, 2015), http://scotsthesaurus.org.

Hüllen, W. *A History of Roget's Thesaurus: Origins, Development, and Design* (Oxford: OUP, 2004).

Hüllen, W. *English Dictionaries, 800–1700: The Topical Tradition* (Oxford: OUP, 1999).

Kay, C. Review of *A Shakespeare Thesaurus*, *International Journal of Lexicography* 9 (1) (1996), 71–76.

Kay, C., and J. Roberts. "Thesaurus." In *The Encyclopedia of Language and Linguistics*, ed. R. E. Asher, 10 vols. (Oxford: Pergamon Press, 1994), 4603–4605.

Kay, C., and M. Alexander. "Diachronic and Synchronic Thesauruses." In *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford: OUP, 2016), 367–380.

Kohli, H. "Transfer Learning and Augmentation for Word Sense Disambiguation." In *Proceedings of the 43rd European Conference on Information Retrieval* (2021), 303–311. https://dx.doi.org/10.1007/978-3-030-72240-1_29.

Lóscio, B. F., et al., eds. "Data on the Web Best Practices: W3C Recommendation 31 January 2017." *W3C* (2017), http://www.w3.org/TR/dwbp/.

Macleod, I., with P. Cairns, C. Macafee, and R. Martin. *The Scots Thesaurus* (Aberdeen: Aberdeen UP, 1990).

McCrae, J. P., A. Rademaker, E. Rudnicka, and F. Bond. "English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology." In

Proceedings of the Multimodal Wordnets Workshop at LREC 2020, eds. T. Declerck, I. Gonzalez-Dios, and G. Rigau (Marseille: European Language Resources Association, 2020), 14–19, http://john.mccr.ae/papers/mccrae2020english.pdf.

McCrae, J. P., and P. Cimiano. "Linghub: A Linked Data Based Portal Supporting the Discovery of Language Resources." In *Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems – SEMANTiCS2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15)co-located with the 11th International Conference on Semantic Systems – SEMANTiCS2015*, CEUR Workshop Proceedings 1481, CEUR-WS.org 2015, eds. A. Filipowska, R. Verborgh, and A. Polleres (Vienna: 2015), 88–91, http://dblp.uni-trier.de/db/conf/i-semantics/semantics2015p.html#McCraeC15.

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

OED Online = Oxford English Dictionary Online (Oxford: OUP, 2000), http://www.oed.com/.

Piao, S. et al. "A Time-sensitive Historical Thesaurus-based Semantic Tagger for Deep Semantic Annotation." *Computer Speech and Language* 46 (2017), 113–135.

Porck, T. *Growing Old among the Anglo-Saxons: The Cultural Conceptualisation of Old Age in Early Medieval England*. Dissertation (Leiden University, 2016).

Roget, P. M. *Thesaurus of English Words and Phrases, Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition* (London: Longman, Brown, Green, and Longmans, 1852).

Sanderson, R., P. Ciccarese, and B. Young, eds. "Web Annotation Data Model: W3C Recommendation 23 February 2017." *W3C* (2017), https://www.w3.org/TR/annotation-model/.

Sapir, E. *Selected Writings of Edward Sapir in Language, Culture and Personality*, ed. D. G. Mendelbaum (Berkely: University of California Press, 1963).

Schandl, T., and A. Blumauer. "PoolParty: SKOS Thesaurus Management Utilizing Linked Data." In *The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30–June 3, 2010, Proceedings, Part II*, Lecture Notes in Computer Science 6089, eds. L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache (Berlin and Heidelberg: Springer, 2010), 421–425, https://dblp.uni-trier.de/db/conf/esws/eswc2010-2.html#SchandlB10.

Scott Jr, E. A. *SPA Design and Architecture: Understanding Single-page Web Applications* (New York: Manning, 2015).

Sharifian, F. "Cultural Linguistics and World Englishes." *World Englishes* 34 (4) (2015), 515–32.

Spevack, M. *A Shakespeare Thesaurus* (Hildesheim: Georg Olms Verlag, 1993).

Sporny, M., D. Longley, G. Kellogg, M. Lanthaler, P. Champin, and N. Lindström. "JSON-LD 1.1: A JSON-based Serialization for Linked Data: W3C Recommendation 16 July 2020." *W3C* (2020), https://www.w3.org/TR/json-ld11/.

Stellato, A., et al. "VocBench 3: A Collaborative Semantic Web Editor for Ontologies, Thesauri and Lexicons." *Semantic Web* 11 (5) (2020), 855–881.

Stolk, S. "*A Thesaurus of Old English* as Linguistic Linked Data: Using OntoLex, SKOS and *lemon-tree* to Bring Topical Thesauri to the Semantic Web." *Proceedings of eLex 2019* (2019), https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_13.pdf.

Stolk, S. "*lemon-tree*: Representing Topical Thesauri on the Semantic Web." In *2nd Conference on Language, Data and Knowledge* (*LDK 2019*), OpenAcces Series in Informatics (OASIcs) 70, eds. M. Eskevich, G. de Melo, C. Fäth, J. P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek, and M. Dojchinovski (Dagstuhl: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019), 16:1–13, https://doi.org/10.4230/OASIcs.LDK.2019.16.

Sula, C. A. and H. V. Hill. "The Early History of Digital Humanities: An Analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004)." *Digital Scholarship in the Humanities* 34 (2019), i190–i206, https://doi.org/10.1093/llc/fqz072.

Suominen, O., H. Ylikotila, S. Pessala, M. Lappalainen, M. Frosterus, J. Tuominen, T. Baker, C. Caracciolo, and A. Retterath. "Publishing SKOS Vocabularies with Skosmos." (2015), https://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf.

Taelman, R., J. Van Herwegen, M. Vander Sande, and R. Verborgh. "Comunica: A Modular SPARQL Query Engine for the Web." In *The Semantic Web – ISWC 2018*, Lecture Notes in Computer Science 11137 (Cham: Springer, 2018), 239–255, https://comunica.github.io/Article-ISWC2018-Resource/.

TOE = Roberts, J., and C. Kay with L. Grundy. *A Thesaurus of Old English* (Glasgow: University of Glasgow, 2017), http://oldenglishthesaurus.arts.gla.ac.uk/.

Tudorache, T., C. Nyulas, N. F. Noy, and M. A. Musen. "WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web." *Semantic Web* 4 (1) (2013), 89–99.

Vea Escarza, R. "Old English Verbs of Envy: Class Membership and Grammatical Behaviour." In *Studies in the History of the English language VIII: Boundaries and Boundary-crossings in the History of English*, eds. P. J. Grund and M. E. Hartman (Berlin: Mouton De Gruyter, 2021), 187–208.

Verborgh, R., S. Wrigley, and R. M. Ballardini. "Decentralizing the Web and the Industrial Internet: An Alternative Solution to Data Control." In *Regulating Industrial Internet through IPR, Data Protection and Competition Law*, eds. R. M. Ballardini, P. Kuoppamäki, and O. Pitkänen (Alphen aan den Rijn: Kluwer Law International, 2019), 47–59.

Wierzbicka, A. *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese* (Oxford: OUP, 1997).

Wilkinson, M. D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (160018) (2016), DOI: 10.1038/sdata.2016.18.

Appendix 1: Example Data Catalogue

221

7

The data catalogue below serves two datasets in Evoke: *TOE*-LLD and 'riddle47'.

```
{
  "@context":
    "https://raw.githubusercontent.com/ssstolk/DCAT-AP/master/
    releases/2.0.0/Draft/dcat-ap_2.0.0.jsonld",
  "$schema": "./catalog.schema.json",
  "@id": "",
  "@type": "Catalog",
  "service": [
    {
      "@id": "http://evoke.ullet.net/platform",
      "@type": "DataService",
      "title": "Evoke platform",
      "identifier": "evoke-platform",
      "endpointURL": "http://142.93.226.251:8081",
      "endpointDescription": "http://evoke.ullet.net/api",
      "landingPage": "http://evoke.ullet.net/app/",
      "mode": "get",
      "servesDataset": [
        "http://oldenglishthesaurus.arts.gla.ac.uk/",
        "https://w3id.org/evoke/set/riddle47"
      ]
    }
  ],
  "dataset": [
    {
      "@id": "http://oldenglishthesaurus.arts.gla.ac.uk/",
      "@type": "Dataset",
      "title": "A Thesaurus of Old English",
      "identifier": "toe",
      "landingPage": "http://evoke.ullet.net/thesaurus/toe/",
      "license": "http://evoke.ullet.net/thesaurus/toe/#license",
      "issued": "2017-05-26",
      "distribution": {
        "accessService": "http://evoke.ullet.net/platform",
        "accessGraph": "http://oldenglishthesaurus.arts.gla
        .ac.uk/",
```

```
        "mediaType": "application/sparql-results+json"
      }
    },
    {
      "@id": "https://w3id.org/evoke/set/riddle47",
      "@type": "Dataset",
      "title": "riddle47",
      "identifier": "riddle47",
      "distribution": {
        "accessService": "http://evoke.ullet.net/platform",
        "accessGraph": "https://w3id.org/evoke/set/riddle47",
        "mediaType": "application/sparql-results+json"
      },
      "requires": [
        "http://oldenglishthesaurus.arts.gla.ac.uk/"
      ]
    }
  ]
}
```

Appendix 2: Example User-Created Annotations

The JSON-LD snippet below contains annotations of two word senses found in *TOE*-LLD, both tagged with the label 'riddle47' through the user interface of Evoke.

```
{
  "@context": [
    "http://www.w3.org/ns/anno.jsonld",
    {"skos": "http://www.w3.org/2004/02/skos/core#",
      "Concept": "skos:Concept",
      "prefLabel": "skos:prefLabel"},
    {"content": {"@reverse": "rdfs:isDefinedBy", "@container":
    "@index"}}
  ],
  "content": {
    "Annotation": [
      {
        "id": "https://w3id.org/evoke/id/annotation/e6081476-b449
        -45d1-bec5-31aad6aad367",
        "type": "Annotation",
        "created": "2020-06-29T17:41:06.351Z",
        "motivation": "commenting",
        "target": "http://oldenglishthesaurus.arts.gla.ac.uk/
        sense/#id=17981",
        "bodyValue": "#riddle47",
        "body": {
```

```
            "id": "https://w3id.org/evoke/id/annotation/e6081476
            -b449-45d1-bec5-31aad6aad367-body",
            "type": "SpecificResource",
            "source": [
            "https://w3id.org/evoke/id/concept/riddle47"
            ],
            "purpose": "tagging"
          }
        },
        {
          "id": "https://w3id.org/evoke/id/annotation/d17922bc-da28
          -4bab-82a3-238776c753ab",
          "type": "Annotation",
          "created": "2020-06-29T17:42:02.782Z",
          "motivation": "commenting",
          "target": "http://oldenglishthesaurus.arts.gla.ac.uk/
          sense/#id=33789",
          "bodyValue": "#riddle47",
          "body": {
            "id": "https://w3id.org/evoke/id/annotation/
            d17922bc-da28-4bab-82a3-238776c753ab-body",
            "type": "SpecificResource",
            "source": [
              "https://w3id.org/evoke/id/concept/riddle47"
            ],
            "purpose": "tagging"
          }
        }
      ],
      "Concept": [
        {
          "id": "https://w3id.org/evoke/id/concept/riddle47",
          "type": "Concept",
          "prefLabel": "riddle47"
        }
      ]
    }
}
```


223

7

Chapter 8

8. Research case studies with *A Thesaurus of Old English* and Evoke

8.1. Introduction

This chapter assesses the usefulness of *A Thesaurus of Old English* (*TOE*) as Linguistic Linked Data, both for research and educational purposes, through a discussion of a number of case studies. These case studies were part of a collaborative project titled 'Exploring Early Medieval English Eloquence' (EEMEE), established specifically for this purpose.[1] The project brought together scholars from universities and lexicographic institutions from across Europe to explore – and elaborate on – the contents of *TOE* using the web application Evoke. Three workshops, organized between 2019 and 2020, facilitated the development and presentation of case studies within the project (see Appendix 8.B). These events informed further refinements to the requirements for Evoke. As described in Chapter 2, functionalities identified and requested for research purposes have been implemented as part of newer iterations of this web application.

Case studies in the EEMEE project approached the thesaurus information from the perspectives of various disciplines: linguistics, literary criticism, history, lexicography, and philology. Participating researchers (and in educational settings, their students) used Evoke to explore the contents of *TOE* alongside additional material. These extended sets of information were fashioned either by linking an existing source to *TOE* data or by using the annotation system within Evoke. Those case studies which were presented at the EEMEE workshop at the 21st International Conference on English Historical Linguistics have been published in a special issue of the international, peer-reviewed journal *Amsterdamer Beiträge zur älteren Germanistik* titled 'Exploring Early Medieval English Eloquence: A Digital Humanities Approach with *A Thesaurus of Old English* and Evoke'.[2] Many of the datasets created in this project have been made available publicly, under an open license (CC-BY-SA), in the DataverseNL data repository and in Evoke (see Appendix 8.A). Their open access availability is intended to encourage future study beyond the scope of the EEMEE project and not restricted to any single software application.

The remainder of this chapter is structured as follows. Section 8.2 provides background information on *TOE* and its relation to Old English language and

[1] EEMEE was supported by the LUCAS Extra Resources Open Call-II Grant 2020, awarded by the Leiden University Centre for the Arts in Society, and the LUCDH Small Grant 2018, awarded by the Leiden University Centre for Digital Humanities.

[2] *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4. Figures included are for illustrative purposes only, stemming from articles or presentations, and are used with permission.

culture. Subsequently, sections 8.3-8.7 present, and reflect on, notable EEMEE case studies on the topics of the history of Old English lexicography, stylistics of Old English writing, diachronic developments of Old English, comparative analyses of Old Germanic languages, and teaching Old English language and culture. The case studies are followed by an evaluation of the usefulness of the Linguistic Linked Data form of *TOE* and the web application Evoke for research purposes in section 8.8.

8.2. *A Thesaurus of Old English* and its relation to Old English language and culture

Originating as a pilot study for University of Glasgow's *Historical Thesaurus* project, *TOE* was quickly recognized as an essential resource in its own right.[3] Since its first publication in 1995, this thesaurus has been the point of departure for many scholars studying aspects of the language and culture of early medieval England. Scholars performing semantic field studies of the language or literary-critical analyses of Old English texts have profited from the ability to look up which Old English words were available to denote particular concepts.[4] The current section explores the relation of this thesaurus to Old English language and culture, examining whose language has been recorded and its relation to the culture of England's early medieval inhabitants. As will become evident, thesauri and related bodies of knowledge can act as "a series of step-by-step doorways into the heart of a national culture".[5]

For their lexis, *TOE* and other dictionaries of Old English have relied on surviving texts of the early medieval language. Spoken between roughly 500 and 1100, the language is found, in runes or a version of the Latin alphabet, in "a rich diversity of records written on parchment, carved in stone and inscribed in jewelry".[6] In total, the corpus of surviving texts, as found in *DOEC*, contains 3,037 texts with over 3 million Old English tokens, i.e., orthographic words rather than lemmas found in dictionaries. The size of this corpus, which is said to equal "almost five times the collected works of Shakespeare",[7] is only a small fraction of the surmised body of texts that must have existed.[8] The variety of the surviving texts notwithstanding, scholars should have an understanding of the writing culture in which these texts were produced in order to interpret their findings based on *TOE* correctly.

A select group in early medieval society could read; an even smaller group

[3] Bremmer, 'Treasure Digging', p. 109; Dance, Review of *TOE1*, p. 312; Görlach, Review of *TOE1*, pp. 398-9.

[4] See, for instance, Bouwer, *Studien zum Wortfeld um* eald *und* niwe *im Altenglischen*; Moriyama, 'Synonyms and Synonymous Expressions in the Old English Semantic Field "Hospitality, Harbouring, and Entertaining"'; Roberts, 'The Old English Vocabulary of Nobility'.

[5] *ScT*, p. ix.

[6] 'About the Dictionary of Old English'.

[7] Ibid.

[8] To illustrate, recent research utilizing unseen species models from ecology estimates 38.6% of narratives from Old and Middle English has survived (Kestemont et al., 'Forgotten books').

could write. Scribes of the surviving Old English texts were often learned monks and their writing reflects their interests: religious topics dominate the corpus.[9] Vivid examples include the many Saints' lives, such as Ælfric's *Lives of Saints*, and adaptations of Biblical books, such as the Old English *Genesis* and *Exodus* in the Junius Manuscript. Any lexicographic work concerned with the Old English lexicon is therefore bound to reflect this writing culture rather than representing the vocabulary of the majority of the populace of the early medieval kingdoms in England. Even so, much can be said about the lexis available to the population. Old English authors shared the vernacular language with the population at large and their writing influenced their society beyond the monasteries, too. Sermons, such as the Old English *Sermo Lupi ad Anglos*, will have been preached in the vernacular to the laity. The content of charters, which recorded transfer of land property, and legal texts will have affected inhabitants in their daily lives. Moreover, thematic domination in the corpus is mitigated in *TOE* through its inclusion of all words and word senses discovered in the corpus, regardless of the frequency with which they occur. Words that are attested infrequently, perhaps only once throughout the corpus, are incorporated alongside those that occur often and across numerous texts.

Analysing the Old English lexicon can offer insights into the culture of the medieval population who spoke it as well as the composition of the historical language. Alaric Hall argues this case convincingly:[10]

> There is [..] a well-established and theoretically justified supposition that language reflects culture [..]. This, as a generalisation, can hardly be denied – if language did not reflect culture then it would be an absurdly ineffectual tool for communication.

That language can convey aspects of the culture using it is by no means a novel notion. Decades ago, Edward Sapir already argued that vocabulary is a "very sensitive index of the culture of a people".[11] Indeed, this notion forms the corner stone of cultural linguistics, in which the relationship between language and culture is explored.[12] *TOE*, through explorations of the Old English lexis, offers such glimpses. Degrees of lexicalization presented by the thesaurus, for instance, are indicative of the salience of concepts and semantic domains.[13] A high degree of lexicalization suggests the need for many nuances in communication by the language community concerned; an altogether lack of words for a concept suggests unimportance or absence. Low degrees of lexicalization for concepts may convey unimportance, too, although they may also be the result of ritualization or other linguistic processes.[14] These and other indicators of cultural aspects

[9] David and Simpson, 'The Middle Ages to ca. 1485: Introduction', p. 5.
[10] Hall, *Elves in Anglo-Saxon England*, p. 13.
[11] *Selected Writings of Edward Sapir in Language, Culture and Personality*, p. 27.
[12] Sharifian, 'Cultural Linguistics and World Englishes', p. 515.
[13] Wierzbicka, *Understanding Cultures through Their Key Words*, pp. 10–11.
[14] In his book chapter 'Old English "Cross" Words', Rolf H. Bremmer Jr identifies six native words in Old English for the cross that bore Jesus Christ, whereas "present-day English today possesses only one current word, 'cross'," a development which could be attributed to ritualisation of the religious vocabulary (p. 231).

encapsulated in the Old English lexicon have been investigated through *TOE* in various case studies of the EEMEE project.

In addition to surviving texts, *TOE* has relied on prior lexicographic work on the language. The process of creating the thesaurus has been described in the prefatory matter of its various editions and, more fully, in a recent article by Jane Roberts, one if its editors.[15] Existing dictionaries of Old English supplied the thesaurus editors with words and word senses, which were copied onto slips of paper and rearranged to form groups of synonyms to be positioned in an overarching, topical structure. Revisions on *TOE* are ongoing, as its current editor updates its content based on the findings of the team behind University of Toronto's *Dictionary of Old English* (*DOE*). The knowledge of the Old English lexis recorded in *TOE* is therefore affected by lexicographic practices and choices during the compilation of these source dictionaries and the resulting thesaurus. The following section will delve more deeply in the history of Old English lexicography and present an EEMEE case study on the subject.

8.3. History of Old English lexicography

Several dictionaries of Old English have been published since the early modern period. The first was William Somner's *Dictionarium Saxonico-Latino-Anglicum*, which appeared in 1659. Almost two centuries later, Joseph Bosworth published his *A Dictionary of the Anglo-Saxon Language* in 1838. Fifty-six years later, John R. Clark Hall issued his *A Concise Anglo-Saxon Dictionary for the Use of Students* in 1894. Both Bosworth's and Clark Hall's dictionary were further developed and expanded, which resulted in publications of supplements and revisions in the late nineteenth and the twentieth century.[16] *TOE*, alongside University of Toronto's *Dictionary of Old English*, is one of the more recent, major lexicographical works of Old English that is still being updated. Developments in Old English lexicography can be scrutinised through the semantic domains available in *TOE*, as demonstrated by the case study below.

8.3.1. Charting lexicographic developments

In her case study, Rachel Fletcher posited that Evoke and the Linguistic Linked Data version of *TOE* enable investigations into the representation of the Old English lexicon throughout the history of Old English lexicography.[17] As mentioned in the previous section, *TOE* presents a filtered image of the early medieval English lexicon: it is based on the surviving texts, interpreted by scholars, and on prior dictionaries. Each lexicographic work reflects its aim,

[15] Roberts, '*A Thesaurus of Old English*: The Pilot Study for the Glasgow *Historical Thesaurus*'.

[16] In 1898, Thomas N. Toller published a revised edition of Bosworth's dictionary under the title 'An Anglo-Saxon Dictionary'. A supplement became available in 1921 and an edition with "enlarged addenda and corrigenda", by the hand of Alistair Campbell, appeared in 1972. Clark Hall's dictionary exists in four editions, of which the last published in 1960.

[17] Rachel Fletcher presented the case study in a paper titled 'Evoke and the History of Old English Lexicography: Preliminary Explorations' at the second EEMEE workshop (see Appendix 8.B).

target audience, the knowledge of the lexis at that particular point in time, and the editorial choices inherent in lexicography. Utilizing Evoke, Fletcher's case study explored how the representation of the Old English lexicon in *TOE* compares to its representation in earlier dictionaries, aiming to "shed light not only on the subjectivity of the *Thesaurus of Old English* (of which users should be aware) but on changing attitudes towards Old English and early mediaeval English culture over the history of scholarship".[18]

The case study investigated five dictionaries alongside *TOE*, including the two principal dictionaries on which *TOE* is based. The five dictionaries are listed below in order of publication date.

- W. Somner, *Dictionarium Saxonico-Latino-Anglicum* (Oxford, 1659).
- E. Lye, *Dictionarium Saxonico et Gothico-Latinum* (London, 1772).
- J. Bosworth, *A Dictionary of the Anglo-Saxon Language* (London, 1838).
- J. Bosworth and T. N. Toller, *An Anglo-Saxon Dictionary Based on the Manuscript Collections of the Late Joseph Bosworth* (London, 1898), *Supplement* by T. N. Toller (Oxford, 1921), with *Enlarged Addenda and Corrigenda* by A. Campbell (Oxford, 1972).
- J. R. Clark Hall, *A Concise Anglo-Saxon Dictionary*, 4th edn, with a supplement by H. D. Meritt (Cambridge, 1960).

Custom labels were created to identify each dictionary ('Som', 'LM', 'Bos', 'BT', 'HM') and used to mark *TOE* word senses in Evoke to reveal which were recorded in the aforementioned dictionaries, thereby facilitating comparisons per semantic domain.

In her case study, Fletcher focused on two *TOE* categories: "13.02.10.01.01 A warrior, fighter" and "14.01.04 A legal right". Figure 8.1 shows an overview of Old English words denoting the former, including the established custom labels as assigned to each. A diachronic visualization of these findings is shown in Figure 8.2. The results foreground that, compared to the current state of Old English lexicography, a disproportionate number of words denoting "A warrior, fighter" were not yet included in earlier dictionaries of Old English. Notably absent in these dictionaries are so-called kennings and other compounds found in poetic diction (e.g., *scildwiga* [lit. "shield warrior"] and *heoruwulf* [lit. "sword wolf"]). Fletcher posited that these findings reflect the historical interest in Old English lexicography, which focused initially on legal and historical sources (e.g., laws, chronicles) and expanded later to poetic texts.[19] In the future, Fletcher concluded, further developments in Old English lexicography may change this picture. Ongoing work on Toronto's *Dictionary of Old English*,

[18] Citation taken from the paper abstract submitted to the workshop.

[19] Fletcher indicated there are various reasons why the poetic lexicon received less attention in early scholarship, including the inaccessibility of some poetic texts (e.g., Vercelli MS was only rediscovered in the nineteenth century) and a higher difficulty in interpreting them (due to their freedom in syntax, lack of Latin parallel texts, and, importantly, use of hapax legomena, words that occur in only a single text throughout the corpus).

is improving scholarly understanding, and surfacing of new material will influence how researchers perceive and study Old English language and culture.

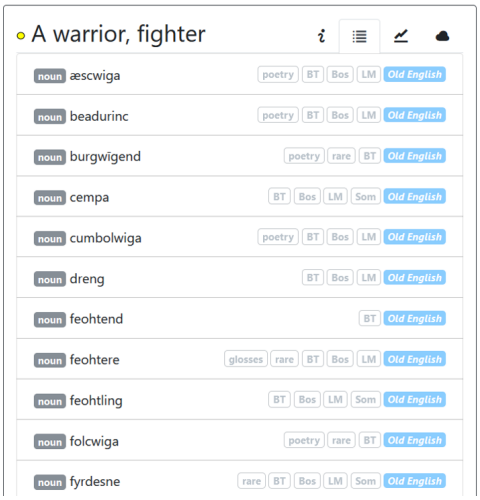


Figure 8.1.: List in Evoke of Old English words denoting "A warrior, fighter", annotated with custom labels indicating dictionaries in which each word has been recorded.

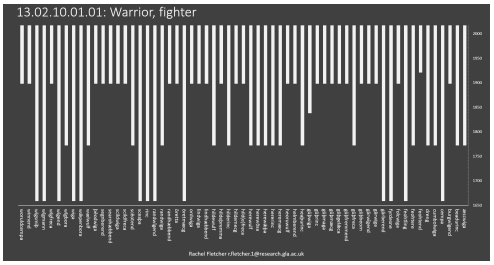


Figure 8.2.: Bar chart of Old English words denoting "A warrior, fighter" and the dates from which they were first recorded in an Old English dictionary.

Fletcher's work foregrounds the value of custom labels for exploring aspects besides stylistics or diachronic developments of a language. Her case study illustrates the use of thesauri for the study of lexicographic practices, although such explorations are not without challenges. Variation in spelling amongst the dictionaries, differences in the granularity of sense definitions, and reinterpretation of words as belonging to another semantic domain can hamper straightforward application of labels and alter what can be deduced from subsequent analysis. Even so, Fletcher's case study demonstrated that *TOE* can provide a semantic framework under which concepts and entire domains can be scrutinized in terms of their lexicographic developments. Annotation functionality in Evoke enables viewing and sharing such findings. The chart in Figure 8.2 indicates a need for further visualizations beyond those currently available in Evoke, and potentially including temporal information, that may inform future development of the web application.

8.4. Stylistics of Old English writing

Historical language thesauri can also be valuable resources for exploring stylistics of specific texts, authors from the historical period, or entire genres. Two EEMEE case studies demonstrated this avenue of research for the Old English context through enrichment of the *TOE* dataset with custom labels. Their subsequent analyses in Evoke have led to new insights into the language use in specific texts

(*Beowulf*, *Andreas*, the *Old English Martyrology*) or by a specific author (Ælfric of Eynsham). These case studies are discussed in separate subsections below.

8.4.1. Onomasiological profiles of Old English texts

In his case study, Thijs Porck demonstrated how Evoke can be used to investigate the lexis used in individual Old English texts.[20] By tagging all words that occur in *Beowulf*, *Andreas*, and the *Old English Martyrology*, Porck expanded the Linguistic Linked Data version of *TOE* with information that allows users to navigate the vocabulary in these specific texts through the semantic hierarchy of the thesaurus. In effect, this process has created three workable prototypes of thesauri, specific to the selected texts, which have been made publicly available in Evoke for anyone to browse and analyse (see Appendix 8.A).

By combining the *TOE* dataset in Evoke with additional information on individual texts, Porck was able to navigate and analyse the vocabulary use in these medieval texts. Figure 8.3, for example, lists Old English words denoting "A man, warrior" and the labels assigned to them. Overviews such as these allow differences in word use between Old English texts to be discerned. The word *ceorl*, for instance, was shown to occur in the poem *Beowulf* but not in *Andreas* or the *Old English Martyrology*, whereas *þegn* occurs in all three texts.

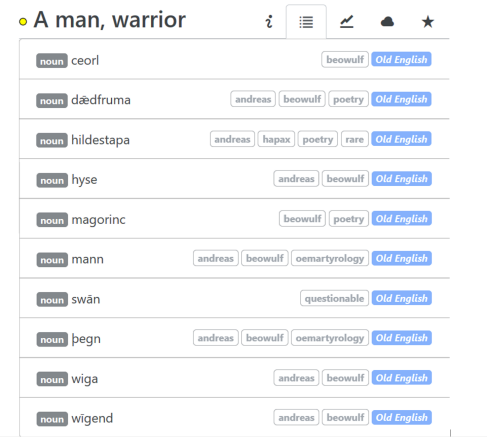


Figure 8.3.: List in Evoke of Old English words (and their labels) denoting "A man, warrior".

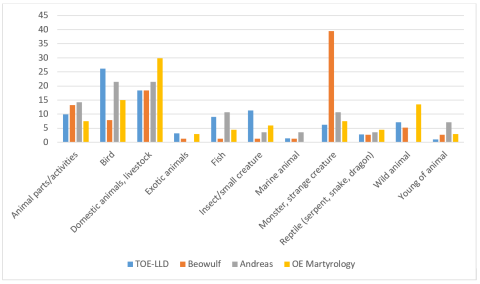


Figure 8.4.: Onomasiological profiles of various Old English texts for the semantic field of "Animal".

Additionally, the combined information in this case study can be used to create semantic fingerprints, or 'onomasiological profiles', of individual Old English texts through the statistics provided by the interface of Evoke. In contrasting such profiles for individual Old English texts, Porck showed how new and distinctive patterns of vocabulary use can be brought to light. To illustrate, Figure 8.4 captures vocabulary use in the semantic field of "Animal" across the entire corpus of Old English and within the three different Old English texts,

[20] Porck, 'Onomasiological Profiles of Old English Texts'.

specifically. Each text, Porck argues, "shows distinctive patterns of vocabulary use that can be related to the roles that animals play in the individual texts".[21] For instance, most of the words used in the *Old English Martyrology* within this semantic field are found in the categories of "Domestic animals, livestock", "Bird" and "Wild animal", which matches scholarly observations on medieval hagiography. As Porck explains, "typically, saints are served by domestic animals and birds, while wild animals either miraculously come to their aid or represent devils in disguise".[22] *Beowulf* displays a vastly different picture. By far, the category of "Monster, strange creature" contains the majority of the words used within this field, which is unsurprising given the prominent role of monsters within the poem.

On a higher, more abstract level, Porck employed onomasiological profiles of the three Old English texts to obtain insights into aspects that include their poetic characterization and degree of ambiguity. On the first aspect, Porck remarks that the "lexicon of *Beowulf* consists of more entries exclusively found in poetic contexts (30.66%) than that of *Andreas* (20.60%), while the lexicon of the prose *Old English Martyrology* is, naturally, devoid of poetic vocabulary".[23] On the degree of ambiguity found in the texts, which is based on the number of polysemous senses attributed to each lexical entry, he notes that texts marked with a higher level of words found solely in poetic contexts (i.e., *Beowulf* and *Andreas*) are more likely to be low in polysemous senses than other texts (i.e., the *Old English Martyrology*) and yield a lower degree of ambiguity.[24]

Porck's use of Evoke and *TOE* exemplifies the value of linking resources and how combining sets of information allows researchers to obtain a picture hitherto unavailable. In fact, an onomasiological analysis of these Old English texts would not go amiss in the introductions to their respective scholarly editions. A case in point is the fourth edition of *Klaeber's Beowulf*, which contains information on, amongst others, the diction, orthographic characteristics, and narrative structure. This material positions the poem in its literary and linguistic context — a positioning which would benefit from insights into the semantic domains present in this text. Contrasts with contemporary texts, across genres, and diachronic developments could, over time, be added to such sections on onomasiological characteristics. Indeed, much remains to be explored. Only a few selected fields have been scrutinized, thus far, using the analysis functionality of Evoke. The textual thesauri fashioned by Porck, publicly available in Evoke, will undoubtedly lead to a deeper understanding of these and other texts in the future.

[21] Ibid., p. 378.
[22] Ibid., p. 379.
[23] Ibid., p. 370.
[24] Ibid., pp. 371-2.

8.4.2. Exploring 'Ælfrician' vocabulary

Amos van Baalen used Evoke in a manner similar to that of Porck.[25] However, rather than creating onomasiological profiles of texts, van Baalen fashioned one for an individual author: Ælfric of Eynsham (c.955x957–c.1010). This abbot was one of the most prominent and prolific writers of the period. His vocabulary – and its influence on the surviving Old English lexis – is therefore important for our understanding of the language and culture of early medieval England. In researching Ælfric's vocabulary, van Baalen employed the *Dictionary of Old English* and prior scholarship to identify and categorize the lexis that is characteristic for the works of Ælfric. The results were used to establish an onomasiological profile in Evoke of the abbot's characteristic vocabulary.

Rather than tagging all words known to have been used by Ælfric, van Baalen first established and categorized those words that had been identified as being representative of 'Ælfrician' vocabulary. Words that are predominantly used by Ælfric (such as *cāsus* 'case' and *forþearle* 'very much, greatly') were marked in Evoke and subsequently analysed in order to create an onomasiological profile, which allowed van Baalen to comment on such aspects as the ambiguity, synonymy, specificity, and semantic distribution of the words characteristic of Ælfric's writing (see Figure 8.5 and 8.6).[26] The use of custom labels in Evoke ensured that the various categories established by van Baalen could be represented that convey the frequency of a word's use by Ælfric, as recorded in *DOEC*, compared to its frequency in surviving Old English texts by other authors.

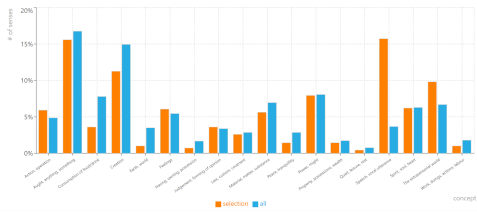


Figure 8.5.: Semantic distribution of Ælfrician vocabulary contrasted with that of all Old English lexis.

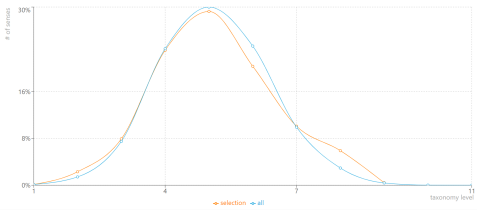


Figure 8.6.: Specificity of Ælfrician vocabulary contrasted with that of all Old English lexis.

Van Baalen's work with Evoke establishes a working method to quantify and pinpoint distinctive authorial preferences within semantic domains. His approach demonstrates that the annotation system of Evoke is an effective means for scholars to add information to a thesaurus and to view as well as query the results. An exciting aspect of this case study is that the categorizations for Ælfrician vocabulary, added by the scholar as labels to the lexis recorded in *TOE*, hint towards future directions in which thesauri and corpora are more

[25] Van Baalen, 'Identifying, Categorising and Exploring "Ælfrician" Vocabulary using the *Dictionary of Old English*, *A Thesaurus of Old English* and Evoke'.

[26] A list of all words found to be typical of Ælfric's writings, listed under the various categories established in this case study, is included as an appendix to the article by van Baalen.

closely integrated. Once word attestations are linked to their corresponding entries in thesauri, onomasiological profiles can consider matters of frequency and predominant usage.[27] Moreover, such an integration would, effectively, allow onomasiological profiles to be generated for authorial preferences, specific texts, and even entire genres.

8.5. Diachronic developments of Old English

In its current, traditional form, *TOE* does not allow researchers to study changes across the Old English period, since, as Richard Dance points out, the thesaurus treats its items as "a single geographically and temporally indistinguishable mass", ignoring diachronic differences between, e.g., Early West Saxon and Late West Saxon.[28] However, Evoke's functionality to add custom labels to words does allow for diachronic investigations by incorporating such distinctions. Information on the origins of words can be added, too, in a similar manner. Jane Roberts asserts that such practice, e.g., flagging Latin loan words or those with an Old Norse origin, "could lead to the examination of loan translation and word formation, dialect vocabulary, etc.; and Evoke provides a welcome platform for such undertakings".[29] The following subsection discusses one approach towards diachronic investigations, undertaken by Khan et al., who have mapped conceptual variation in *TOE* through the web application Evoke.

8.5.1. Mapping conceptual variation and diachronic changes

The case study by Khan et al. explored how Evoke may be used to visualize, navigate, and investigate conceptual mappings within the Old English vocabulary of SHAME.[30] The authors, after having established a list of 28 words and expressions for SHAME based on *TOE* and other lexicographic resources, charted conceptual variation in this vocabulary through two classifications. Firstly, each word or expression was classified as a metaphorical mapping. For instance, the verb *ārēodian* ("to blush, be ashamed", lit. "to turn red") belongs to the mapping EMOTION IS REDNESS IN THE FACE. Secondly, the development of each word or expression for SHAME was classified into one of four diachronic scenarios based on whether or not a sense shift must have taken place between Proto-Germanic and Old English.

The results of their classifications were captured in Evoke using its annotation system. The Old English SHAME words and expressions in *TOE* were tagged with their metaphor mapping (including their source and target) and the diachronic scenario to which they belong (see Figure 8.7). Khan et al. showed that these annotations facilitate the discoverability and further analysis of conceptual

[27] Standardizing the modelling of frequency and attestations in Linguistic Linked Data is currently being pursued (Chiarcos et al., 'Modelling Frequency and Attestations for OntoLex-Lemon').

[28] Dance, Review of *TOE1* (p. 313).

[29] Roberts, '*A Thesaurus of Old English*: The Pilot Study for the Glasgow Historical Thesaurus', p. 312.

[30] Khan et al., 'Mapping Conceptual Variation through *A Thesaurus of Old English* and Evoke'.

variation in Old English figurative language. The web application generated overviews of entire mappings or specific diachronic developments on demand (see Figure 8.8). The authors indicated that they intend to explore these research avenues further in future work.

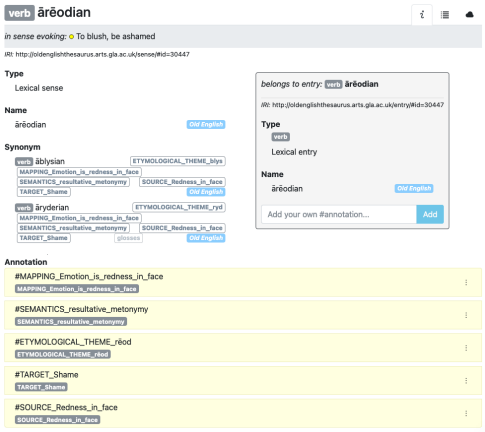


Figure 8.7.: A lexical sense annotated in Evoke with semantic mappings.

Figure 8.8.: Overview in Evoke of senses with the metaphorical use EMOTION IS NAKEDNESS.

As Khan et al. have demonstrated, thesauri are valuable resources for research on metaphors. Indeed, the value of these lexicographic resources for identifying and charting metaphorical relations between semantic fields has been reported in other research projects, too.[31] Additionally, Khan et al. revealed advantages to their having used Evoke. The annotation system of the application and the overviews generated offered the means to interact more closely with the source material. A thesaurus edition can thus not only be a source of information but also act as a canvas for information produced. Such interaction with lexicographic resources may well facilitate feedback loops to provide the original editors with useful comments and suggestions that can be incorporated in revisions.

8.6. Comparative analyses of Old Germanic languages

Two further EEMEE case studies demonstrated the value of thesauri for comparative analyses of related languages. The first, by Rita van de Poel and Sander Stolk, linked Old Frisian lexis with that of Old English captured in the *TOE* dataset. The second case study, by Katrien Depuydt and Jesse de Does, connected Old Dutch to Old English. The resulting data connections provide an avenue of research for contrasting kindred languages through the overarching, onomasiological framework of the thesaurus. Thus, commonalities and differences between languages can emerge based on their representation of specific semantic fields.

[31] *Mapping English Metaphor through Time.*

8.6.1. Old Frisian and Old English KINSHIP

The case study by Rita van de Poel and Sander Stolk used Evoke to link Old Frisian to Old English words captured by the *TOE* dataset.[32] In their work, the authors focused on the semantic field of KINSHIP in the two related languages. In connecting Old Frisian lexis, drawn from a dictionary of Old Frisian, to the overarching structure of *TOE*, the researchers created a dataset that positions Old English and Old Frisian lexis in the same semantic framework. The connected resources were shared and analysed using Evoke (see Figure 8.9).

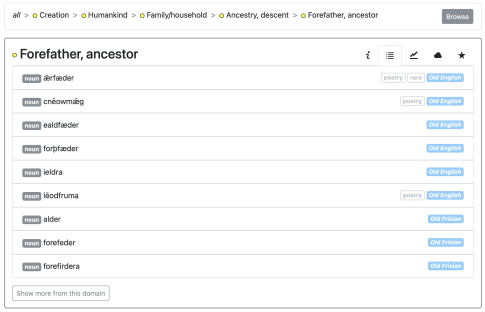


Figure 8.9.: An overview in Evoke of Old English and Old Frisian words denoting "Forefather, ancestor".

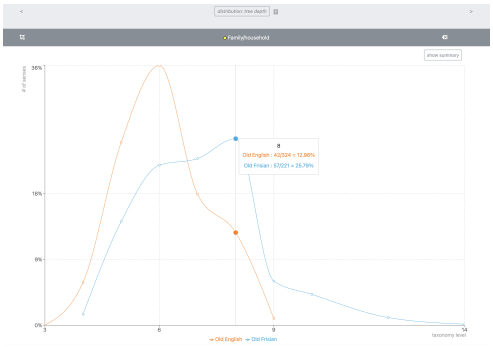


Figure 8.10.: Levels of specificity for Old Frisian and Old English KINSHIP.

Van de Poel and Stolk demonstrated how Evoke facilitates comparative analyses between the two languages and argue that onomasiological analyses (such as shown in Figure 8.10) can uncover new linguistic and cultural insights. Their case study revealed how the two language communities conceptualized aspects relating to family, ancestry, descent, and adoption.[33] Additionally, van de Poel and Stolk signaled that new thesauri can be created through adoption (and expansion) of the taxonomy of an existing thesaurus of a kindred language — one semantic field at a time. Evoke and the Linguistic Linked Data paradigms support such endeavours, as they allow researchers to access and extend *TOE* data whilst abiding by the license of the thesaurus, which restricts users from extracting or downloading substantial portions of the dataset.

In addition to presenting their findings, van de Poel and Stolk shed light on the major hurdles and limitations in connecting material from two lexicographic resources and analysing the results.[34] Comparisons such as theirs, they note, are influenced by differences in lexicographic practices and editorial choices. The editors of *TOE* and those of the dictionary of Old Frisian will have had, or worked with, different approaches towards the granularity of word senses to include.[35]

[32] Van de Poel and Stolk, 'A Case of Kinship'. The article in its entirety is included in this dissertation and follows the current chapter.

[33] To illustrate, "Adoption" lacks Old Frisian senses entirely, whereas the concept of "Siblings" is lexicalized in this language but not in Old English (Ibid., p. 475).

[34] Ibid., p. 477.

[35] The practice of maintaining only generic word senses is not uncommonly referred to as

Additionally, positioning the Old Frisian lexis in *TOE* categories was complicated by the use of different languages for the sense definitions (i.e., English in *TOE* versus German for the Old Frisian lexis). Nevertheless, the researchers found ways to overcome these issues and succeeded in aligning the two lexicographic resources, thereby adopting the overarching structure of *TOE* as a comparative framework for the two related historical languages. Thus, the scholars could avoid fashioning an onomasiological structure for Old Frisian from the ground up and subsequently, prior to analysis, attempting to reconcile it with that of *TOE*. The resulting, combined dataset allows researchers of both Old Frisian and Old English to adopt the same structure and to identify and discuss notable differences in conceptualizations between the kindred languages through shared semantic domains.

8.6.2. Linking Old Dutch to Old English

Katrien Depuydt and Jesse de Does took a more technical route than Rita van de Poel and Sander Stolk towards linking lexis from another language to *TOE*.[36] In their case study, they reflected on their experiments with manual and automatic techniques in order to establish connections between entries from the *Dictionary of Old Dutch* and the *TOE* dataset available in Evoke. In doing so, they investigated whether the existing macrostructure of *TOE* can be reused to create a thesaurus of Old Dutch.

Their research compared two approaches to linking the two lexicographic resources. The first approach was a manual one, in which the tool Lex'it was used to view entries of Old Dutch words and record appropriate categories from the *TOE* hierarchy for their registered senses (see Figure 8.11). Additional matches were obtained by drawing on cognate information. Cognate relations of Old Dutch entries were traversed to Old Frisian ones that had already been linked to *TOE* categories (based on the work by van de Poel and Stolk described in section 8.6.1). The second approach explored by the authors was semi-automated linking, where five different matching algorithms were contrasted and evaluated. Depuydt and de Does concluded that further directions for linking such lexicographical resources may be found in improving automated linking techniques, crowdsourcing part of the process, and improving visualization of candidate links for verification purposes.

The case study demonstrated two key advantages of the Linguistic Linked Data form in which *TOE* has been made available. Firstly, connections between Old Dutch lexis and *TOE* categories could be established and stored separately from the *TOE* dataset. This possibility ensures that researchers other than the editors of *TOE* can extend the semantic framework formed by the thesaurus, even when they do not have direct access to the raw data of the thesaurus or to the database in which it is stored. Secondly, information linked in previous

lumping, whereas distinguishing fine-grained senses is termed splitting (see Fontenelle, 'Bilingual Dictionaries', pp. 45-6; Kay and Alexander, 'Diachronic and Synchronic Thesauruses', p. 370).

[36] Depuydt and de Does, 'Linking the *Dictionary of Old Dutch* to *A Thesaurus of Old English*'.

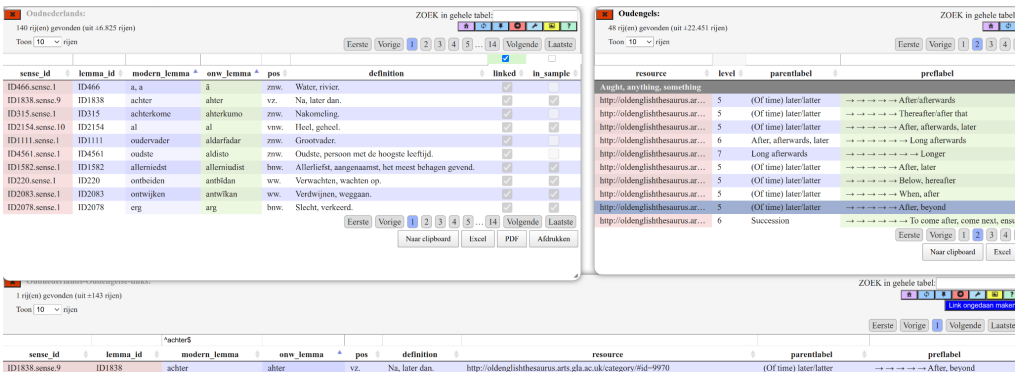


Figure 8.11.: The Lex'it-based linking tool; linking the entry for Old Dutch *ahter* to the *TOE* category "05.11.07.04.02|13 (Of time) later/latter: After, beyond".

research could be queried alongside the original thesaurus material and, thus, drawn on for analyses. In this case, the Old Frisian lexis for KINSHIP, positioned in thesaurus categories by van de Poel and Stolk (see 8.6.1), could be combined with cognate information and leveraged for positioning Old Dutch lexis within the onomasiological structure of the thesaurus.

A dataset containing the established connections between the Old Dutch dictionary material and *TOE* has been made available on a webpage dedicated to the case study.[37] The availability of this dataset enables researchers to perform comparative analyses of Old Dutch and Old English. In their article, Depuydt and de Does do not yet include such an analysis. Their insights on the reuse of an existing onomasiological framework – i.e., that of *TOE* – are therefore limited to the difficulties encountered in creating suitable connections. The language-specific nature of *TOE*, built as it was for "fitting the intricacies of Old English lexis", is one of the reasons the authors put forward for the complexity of matching items from other languages.[38] Additionally, an unfamiliarity with the structure of the thesaurus can impede locating the categories that are of interest. Indeed, Depuydt and de Does indicate that manually assigning Old Dutch lexis to suitable categories within the onomasiological structure of *TOE* was hampered for this reason.[39] Therefore, studies that are sizable in scope may warrant gaining an intimacy with the topical system of a thesaurus before adopting it for other languages and comparative analyses.

8.7. Teaching Old English language and culture

The last aspect of historical language thesauri explored in EEMEE was their didactic potential. Kees Dekker, at University of Groningen, and Thijs Porck and Sander Stolk, at Leiden University, employed *TOE* and Evoke to introduce students to notable aspects of Old English language and culture. In both cases,

[37] https://ivdnt.org/corpora-lexica/diamant/onw-toe-linking/

[38] Depuydt and de Does, 'Linking the *Dictionary of Old Dutch* to *A Thesaurus of Old English*', p. 509.

[39] Ibid., p. 510.

Bachelor students were asked to discern nuances within semantic domains and theorize as to how these might reflect the culture of the Anglo-Saxons. The outcome of the experiment was that students were shown to benefit from extensive use of these resources throughout a course, as is the case with University of Groningen, but also from more concentrated use, such as in workshops at Leiden University. Experiences of the two universities are treated below in separate sections.

8.7.1. The Old English classroom at University of Groningen

In his case study, Kees Dekker employed *TOE* and Evoke in a classroom context through a series of increasingly complicated search assignments.[40] These assignments, intended to interest, challenge, and instruct students, were preceded by an introduction into lexical semantics to inform students of such notions as semantic fields, hyponymy, and synonymy — notions at the core of *TOE* and other thesauri. During a seven-week module, students received assignments on *TOE* and Evoke that were related to the Old English texts they were asked to read and translate. The assignments ranged from, at the start of the module, navigating the thesaurus to take note of the lexis and distinctions available for bodies of water (see Figure 8.12) to, at a more advanced stage, questioning the notion of 'foreign' in Old English (see Figure 8.13) by discerning what the early medieval speakers of this language considered foreign and what connotations their words for this concept had. Through his discussion of both the assignments for students and their responses, Dekker demonstrated that Evoke is a valuable addition to a teacher's tool set, as it invites students and researchers alike to explore Old English in novel and interesting ways.

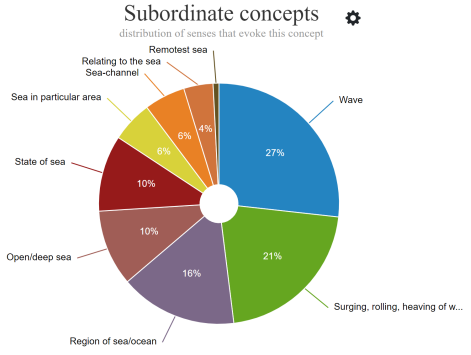


Figure 8.12.: Pie chart in Evoke showing the distribution across the subordinate concepts of the *TOE* category "Sea/ocean".

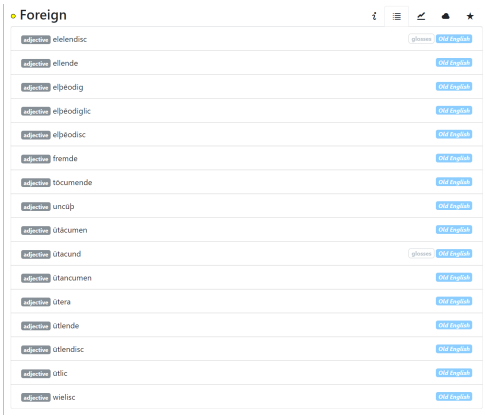


Figure 8.13.: List of Old English words in Evoke denoting "foreign".

Although the requirements for Evoke were formulated to support research first and foremost,[41] Dekker's article demonstrates that the design of the web

[40] Dekker, 'Evoke and *A Thesaurus of Old English* in the Old English classroom'.
[41] Stolk, 'Evoke'.

application enables students, too, to explore a thesaurus and engage with the lexical material presented. The functionality to navigate and view thesaurus material supports users without expertise or prior knowledge of Linguistic Linked Data or thesauri. Moreover, the many visualizations included in Evoke allow students to engage with the material in a playful manner: Next to pie charts, which provide statistics beneficial to research, wordcloud visualizations of entire semantic fields (see Figure 8.14) have led to students calling Evoke "fun" as well as "useful" (see Appendix 8.C, Table 8.C.6).

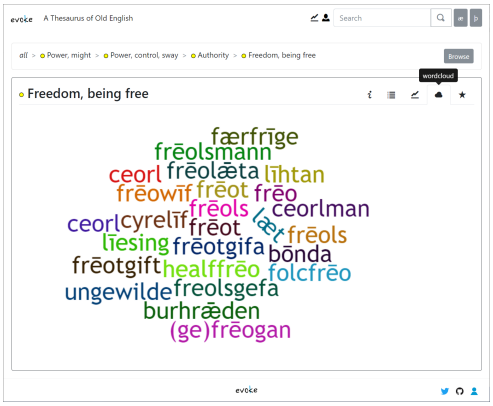


Figure 8.14.: Wordcloud in Evoke for "Freedom, being free".

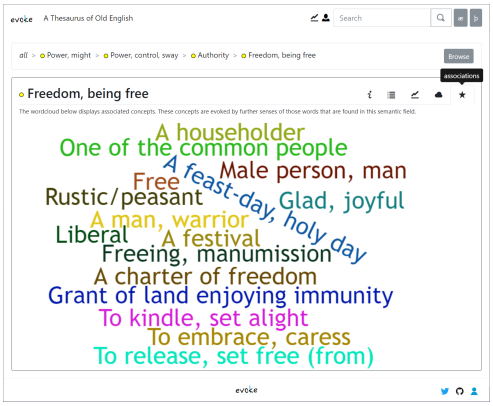


Figure 8.15.: Associations in Evoke for "Freedom, being free".

For future use of Evoke in educational settings, Dekker points out that, although already valuable, the web application could be improved by incorporating "a repository of questions and assignments shared by instructors interested in the use of Evoke for curricular and research purposes".[42] Such an addition is indeed already planned. Existing teaching material will be incorporated in the Evoke website. Preliminary work has been done towards its realisation. 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*.

8.7.2. Workshops on Old English resources at Leiden University

Dekker's findings on the use of Evoke in education are reinforced by those at Leiden University, where the web application has been used in a classroom setting, annually, since 2018. Third-year Bachelor students perform a number of assignments in Evoke during a 2-hour workshop on digital tools and resources for studying Old English, led by Thijs Porck and Sander Stolk. The first of these workshops, which took place on 13 November 2018, included an evaluation by students of the application (see Appendix 8.C). Students considered navigation in Evoke to be "clear" and "easy" to understand. Moreover, views and analyses offered by the application enabled them to obtain results to the following questions, amongst others, in an "efficient" manner:

[42] Dekker, 'Evoke and A Thesaurus of Old English in the Old English classroom', p. 526.

> How many words does Old English have for Hell and its subcategories? How about for Heaven?

The statistics tab in Evoke provides exactly such figures at a glance, avoiding the need for students and researchers to manually count the words listed under various categories.

A recurring element in assignments for students at both Leiden University and University of Groningen has been to establish connotations of words or semantic fields in their entirety. A case in point is the following illustrative question:

> The *TOE* lists several words for old age, including *gelefed* 'old', *har* 'old' and *frod* 'old'. Find out what else these words mean. Search for "gelefed", "har" and "frod". What, according to your findings, did the Anglo-Saxons associate with old age?

Students viewed each word listed separately to ascertain their various senses and, thus, indications of connotations of words used in a certain meaning. During the second EEMEE workshop, it became clear that obtaining connotations of an entire semantic field is deemed valuable for education purposes but, with the then available functionality of Evoke, still required students to engage with content in this piecemeal fashion. A new visualization was therefore introduced. The 'associations' wordcloud, available since version 1.4.1 of Evoke, provides an overview of all senses in a semantic field related through polysemy (see Figure 8.15). In short, feedback on the use of Evoke and *TOE* in education has been valuable in informing the development of Evoke.

8.8. Discussion

This section discusses the usefulness of the digital resources at the heart of the case studies presented above: the Linguistic Linked Data form of *TOE* and the web application Evoke. Both resources have been developed specifically for facilitating research, which raises the question in what manner, if any, they have enabled new lines of scholarly enquiry compared to existing editions of historical language thesauri.

8.8.1. *A Thesaurus of Old English* as Linguistic Linked Data

The Linguistic Linked Data form of *TOE* (or *TOE*-LLD), as made available in Evoke, enabled researchers to extend the original dataset of the thesaurus with information salient to their inquiries. In contrast to many existing editions of historical language thesauri of Scots and English, this digital form includes unique identifiers (IRIs) for lexical items as well as categories.[43] These identifiers

[43] To illustrate, *HTE* does not issue IRIs for a specific lexical sense. The search hit *strength* as found at category "02.02.06.01|06 (n.) Strength of evidence", for example, refers to the following subpath on the *HTE* website: /category/?id=123117&qsearch=strength. This web address indicates that category 123117 is to be shown, i.e., "Strength of evidence", whilst highlighting any text that matches the string 'strength'. The web address is therefore not an exact reference to the specific item requested, but a visual aid that potentially foregrounds

have facilitated explicit references and, in the case of *TOE*-LLD, are unaffected by potential modifications to the spelling of the headword. Being able to reference, view, and link information to lexical items has proven valuable for research in the case studies described above. Moreover, the inclusion of lexical entries in *TOE*-LLD allows researchers to mark items of interest more broadly than per individual lexical sense (cf. the case study by Thijs Porck). These entries, deduced automatically from contextual information available in the original *TOE* dataset, may still need manual revision in places: some homonyms have been lumped together and senses of lexemes remain ungrouped when their headwords use inconsistent spelling or grammatical forms.[44] To eliminate the inherent caveats of such an automated process, thesaurus editors may wish to maintain lexical entries explicitly in order to facilitate these research needs.

Another important advantage in adopting IRIs is that their use simplified extending *TOE* content without having access to the raw data of the thesaurus. Scholars were able to capture knowledge and, in their data, include explicit references to information found in *TOE*-LLD. This capability is advantageous for research, since the license of *TOE* prohibits users from downloading large amounts of thesaurus data.[45] Licenses that restrict access in such a manner are not uncommon for lexicographic resources available on the Web.[46] The *TOE*-LLD version thus lowers the barrier for scholars to engage with the available material.

In addition to the usefulness of IRIs for linking data, even when restrictive licenses are in place, Christian Chiarcos et al. claim two further benefits to use of Linguistic Linked Data.[47] The first is an increased level of interoperability owing to the reuse of established data vocabularies (i.e., Lemon-OntoLex and SKOS).[48] Improved interoperability manifests itself in the shape of software other than Evoke being able to operate on the same data and/or data form. Examples of tools used in the various case studies include custom alignment tools, Microsoft Excel, and Lex'it.[49] Compatible software beyond those employed in the case studies exists, too. Cases in point are VocBench 3 and LexO, with which users can fashion and maintain such data.[50] Current endeavours with Linguistic Linked Data suggest an uptake of this digital form and tools that can operate on it over the next few years. These include the H2020 projects ELEXIS (2018–22), Prêt-à-LLOD (2019–22), and the COST Action NexusLinguarum (2019–23).[51]

---

other entries, too, that include the aforementioned string. For further information on IRIs, see Chapter 3.

[44] See Porck, 'Onomasiological Profiles of Old English Texts', p.366.

[45] See *TOE4*, section 'Copyright and Conditions of Use'.

[46] See, for instance, *Oxford English Dictionary Online*, *HTE*, and the historical Dutch dictionaries part of the *Geïntegreerde Taalbank*.

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

[48] See Chapter 3.

[49] These tools are described in Chapter 6, section 4.3.2, 'Linking Data'. The 'List of source code' in the back matter of the dissertation includes relevant links to code developed for software applications and data transformations, which includes the custom alignment tools.

[50] Stellato et al., 'VocBench 3'; Bellandi and Giovannetti, 'Involving Lexicographers in the LLOD Cloud with LexO'.

[51] ELEXIS: https://cordis.europa.eu/project/id/731015, 2018–2022; Prêt-à-LLOD: https://

Moreover, the form should allow thesauri other than *TOE* (and its extensions) to be made accessible in Evoke in the future without reconfiguration of the database or alterations to the source code to accommodate the new thesaurus.

The second advantage of LLD over traditional Web-based storage, Chiarcos et al., argue, is the ability to merge distinct datasets and thereby obtain a validly formatted, combined set of information. The characteristic of LLD in merging data, thus, allows for complementary data to be queried in unison or different perspectives to be related. The case studies above displayed the use of this functionality in complementing *TOE* with Old Frisian lexis and cultural concepts (see van de Poel and Stolk) or Old Dutch lexis (see Depuydt and De Does). Similarly, annotations and labels can exist in a dataset separate from *TOE* and nonetheless be viewed and analysed cohesively (see Porck; van Baalen). Achieving these ends with the traditional database of *TOE* would be possible too, albeit with a higher effort. Modifications to the database structure can add support for other languages, annotations, and custom labels, but such changes will need to be integrated into the source code underlying the presentation. In short, the proclaimed advantages of LLD have facilitated the case studies described in this chapter and may benefit future studies in a similar manner.

The advantages of the LLD form of *TOE* notwithstanding, use of this resource for research is not necessarily unproblematic. The EEMEE case studies revealed a number of hurdles in connecting additional data to *TOE*. Firstly, differences can be expected between the lexicographic choices that underlie the thesaurus and additional data stemming from another lexicographic resource.[52] Secondly, *TOE* contains conceptualisations, in the form of categories positioned in its topical system, which may work well for the Old English lexis for which the thesaurus has been built from the ground up, but may not be wholly adequate for organizing material from other languages: Concepts required for the additional material may be absent in the existing onomasiological framework (e.g., "Siblings" in the case of Old Frisian) or structured differently from what the additional material would find most useful.[53] Lastly, even when an appropriate category is available in which to position lexis, locating that category is not always straightforward when one is not intimate with the thesaurus structure.[54] Although automation of aligning such resources may help speed up the process, existing techniques thereto are not perfect.[55] These challenges are, of course, not brought about by the use of LLD as digital form of historical language thesauri, but are nonetheless relevant for employing *TOE*-LLD for research.

cordis.europa.eu/project/id/825182, 2019–2022; NexusLingarum: https://www.cost.eu/actions/CA18209,2019\T1\textendash2023.

[52] E.g., Porck, 'Onomasiological Profiles', pp. 364-9; van de Poel and Stolk, 'A Case of Kinship', pp. 477-8.

[53] Ibid.

[54] Depuydt and de Does, 'Linking the Dictionary of Old Dutch to A Thesaurus of Old English', pp. 509-10.

[55] Ibid., p. 508.

8.8.2. Evoke and its functionalities

Chapter 2 identified functionalities for editions of historical language thesauri that would benefit research. Some of these functionalities – i.e., mostly those surrounding navigation and resource views – could already be found in existing editions of thesauri of Scots and English. In contrast, nearly all functionality geared towards extension of original thesaurus content, statistical analyses, and data management is novel.[56] All five groups of functionality were employed in the case studies reviewed in this chapter. The prospect of novel features becoming available to them encouraged researchers to formulate new lines of questions, ones that hinged on either the ability to insert custom labels or the means to position lexis from languages other than Old English in *TOE*. A suite of advanced analysis options, which utilize the topical structure of the thesaurus, provided researchers with insights on their own information complementing the original thesaurus content, effectively creating onomasiological profiles.

Not all functionality listed in Chapter 2 is currently available in Evoke. Most notably, the ability to filter words when browsing, based on user-specified word features, has not been implemented yet. Such functionality would bring about focused views of subthesauri, including the 'Beowulf Thesaurus', 'Ælfrician Vocabulary' and 'Old Frisian: Kinship'. Users of these datasets (created in the case studies by Porck, van Baalen, and van de Poel and Stolk) would benefit from not having to inspect the characteristics of words found at a specific category (e.g., their labels or language) in order to determine whether they belong to the subthesaurus in question.

Additionally, exploration of thesaurus content could be enhanced further. To illustrate, a top 10 of categories containing lexical items conforming to a selection of features salient to the user would facilitate localizing points of interest in the thesaurus structure. Similarly, graphs generated by analyses in Evoke could be supplemented with samples of words or word senses that constitute the figures presented in that graph, thereby making results more tangible and assisting the user in identifying inclusion of any undesired material that warrants refinement of the initial query.

Beyond improvements for exploration, the case study by van Baalen demonstrates another future direction for work on thesaurus editions. Vocabulary characteristic of Ælfric could be identified with higher precision by linking attestations of lexical items in corpora and determining their frequency of occurrence. Such insights would, of course, not be advantageous solely in the context of Ælfric's writings. A selection of single texts or entire genres would be made possible, too. Previous research has explored how such links between *TOE* and the corpus *DOEC* can be represented as Linguistic Linked Data.[57] With the availability of such links, analyses that traverse them and obtain relevant figures would be possible to realize.

The set of functionality identified in Chapter 2, alongside the features put

[56] Only functionality for basic analyses was provided by some editions of historical language thesauri (see Chapter 2).

[57] Chiarcos et al., 'Modelling Frequency and Attestations for OntoLex-Lemon', pp. 7-8.

forward in the previous two paragraphs, is by no means exhaustive. Additional workshops and case studies, increasing the diversity of disciplines and interests represented, are likely to elicit further functionality desired for research and education. Digital Humanities analyses of historical sources could, for example, be given more attention in future research programmes involving thesauri such as *TOE*. Through iterations of software use and development, tools available to researchers can improve incrementally and expedite expansion of our knowledge of historical languages and cultures.

8.9. Conclusion

This chapter has provided an overview of case studies, in research and education, that take advantage of the Linguistic Linked Data form of *A Thesaurus of Old English* and the web application Evoke. The combination of these resources facilitates enrichment of the original *TOE* dataset with information regarding, amongst others, lexicographical history, stylistics, diachronic developments, and kindred languages. Onomasiological analyses, possible through the user interface of Evoke, have led to new insights into Old English language and culture for both students and researchers. This is not to say that all research possible with *TOE* would benefit equally from the Linguistic Linked Data format and the functionality identified in Chapter 2. Smaller scale studies that focus on a single category, for instance, may not directly benefit from the availability of features that perform advanced analyses over the entire thesaurus. Functionality already covered by existing software may necessitate resorting to other, supported data formats. More importantly, the case studies and the resulting materials discussed in this chapter show that the development of digital research tools for historical language thesauri can be a powerful impetus for future research. The new resources at the centre of these studies – *TOE*-LLD and Evoke – have introduced new instruments, embraced by researchers and students participating in EEMEE, that should prove useful beyond the scope of this research project.

References

'About the Dictionary of Old English', *The Dictionary of Old English*, ed. C. Monahan (Toronto, 2022). https://doe.artsci.utoronto.ca/?page_id=11.

Bellandi, A. and E. Giovannetti, 'Involving Lexicographers in the LLOD Cloud with LexO, an Easy-to-Use Editor of Lemon Lexical Resources', Proceedings of the 7th Workshop on Linked Data in Linguistics, Marseille, 22-3 June 2020, pp. 70–4. https://aclanthology.org/2020.ldl-1.10.pdf.

Benedek, J. and T. Miner, 'Measuring Desirability: New Methods for Evaluating Desirability in a Usability Lab Setting', Proceedings of Usability Professionals Association 2002, Orlando, 8-12 July 2002, pp. 8–12.

Bosworth, J., *A Dictionary of the Anglo-Saxon Language* (London, 1838).

Bosworth, J. and T. N. Toller, *An Anglo-Saxon Dictionary Based on the*

Manuscript Collections of the Late Joseph Bosworth (London, 1898), *Supplement* by T. N. Toller (Oxford, 1921), with *Enlarged Addenda and Corrigenda* by A. Campbell (Oxford, 1972).

Bouwer, H., *Studien zum Wortfeld um* eald *und* niwe *im Altenglischen* (Heidelberg, 2004).

Bremmer Jr, R. H., 'Treasure Digging in the Old English Lexicon', Review of *TOE2*, *NOWELE* 40 (2002), 109–14.

Bremmer Jr, R. H., 'Old English "Cross" Words', in *Cross and Cruciform in the Anglo-Saxon World: Studies to Honor the Memory of Timothy Reuter*, eds. S. Larratt Keefer, K. L. Jolly, and C. E. Karkov, Medieval European Studies XI (Morgantown, 2010), 204-32.

Chiarcos, C. et al., 'Towards Open Data for Linguistics: Lexical Linked Data', 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.pdf.

Clark Hall, J. R., *A Concise Anglo-Saxon Dictionary for the Use of Students* (London, 1894).

Clark Hall, J. R., *A Concise Anglo-Saxon Dictionary*, 4th edn, with a supplement by H. D. Meritt (Cambridge, 1960).

Dance, R., Review of *TOE1*, *Medium Ævum* 66.2 (1997), 312–13.

David, A. and J. Simpson, 'The Middle Ages to ca. 1485: Introduction', *Norton Anthology of English Literature*, ed. S. Greenblatt, 2 vols, 8th edn (London, 2006), I, 1–21.

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.

DOE = Dictionary of Old English: A to I Online, eds. A. Cameron et al. (Toronto, 2018). https://tapor.library.utoronto.ca/doe/.

DOEC = Dictionary of Old English Web Corpus, eds. A. diPaolo Healey et al. (Toronto, 2009). https://tapor.library.utoronto.ca/doecorpus/.

'Europeana Impact Playbook', *Europeana*. https://pro.europeana.eu/page/europeana-impact-playbook.

Fontenelle, T., 'Bilignual Dictionaries: History and Development; Current Issues', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 44–61.

Geïntegreerde Taalbank (Leiden, 2018). https://gtb.ivdnt.org.

Görlach, M., Review of *TOE1*, *Anglia* 116.3 (1998), 398–401.

Hall, A., *Elves in Anglo-Saxon England: Matters of Belief, Health, Gender and*

Identity, Anglo-Saxon Studies 8 (Woodbridge, 2007).

HTE = The Historical Thesaurus of English, eds. C. Kay et al. (Glasgow, 2014). http://historicalthesaurus.arts.gla.ac.uk.

Kay, C. and M. Alexander, 'Diachronic and Synchronic Thesauruses', in *The Oxford Handbook of Lexicography*, ed. P. Durkin (Oxford, 2016), pp. 367–80.

Kestemont, M. et al., 'Forgotten Books: The Application of Unseen Species Models to the Survival of Culture', *Science* 375.6582 (2022), 765–9.

Khan, F. et al, 'Mapping Conceptual Variation through *A Thesaurus of Old English* and Evoke: Towards a Topical Thesaurus of Old English Emotional Expressions', *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), 442–56. doi: 10.1163/18756719-12340238.

Learning with the Online Thesaurus of Old English (TOE), eds. C. Hough and C. Kay (Glasgow, 2017). http://oldenglishteaching.arts.gla.ac.uk/.

Lye, E., *Dictionarium Saxonico et Gothico-Latinum* (London, 1772).

Mapping English Metaphor through Time, eds. W. Anderson, E. Bramwell, and C. Hough, (Oxford, 2016).

Moriyama, H., 'Synonyms and Synonymous Expressions in the Old English Semantic Field "Hospitality, Harbouring, and Entertaining"', *Waseda Global Forum* 7 (2010), 163-92.

Oxford English Dictionary Online (Oxford, 2000). http://www.oed.com/.

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.

Roberts, J., 'The Old English Vocabulary of Nobility', in *Nobles and Nobility*, ed. A. J. Duggan (Cambridge, 2000), pp. 69-84.

Roberts, J., '*A Thesaurus of Old English*: The Pilot Study for the Glasgow *Historical Thesaurus*', *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), 297–317. doi: 10.1163/18756719-12340234.

ScT = The Scots Thesaurus, eds. I. Macleod et al. (Aberdeen, 1990).

Selected Writings of Edward Sapir in Language, Culture and Personality, ed. D. G. Mendelbaum (Berkeley, 1963).

Sharifian, F., 'Cultural Linguistics and World Englishes', *World Englishes* 34.4 (2015), 515–32.

Somner, W., *Dictionarium Saxonico-Latino-Anglicum* (Oxford, 1659).

Stellato, A., et al., 'VocBench 3: A Collaborative Semantic Web Editor for Ontologies, Thesauri and Lexicons', *Semantic Web* 11.5 (2020), 855–81.

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), 318–58. doi: 10.1163/18756719-12340235.

TOE = A Thesaurus of Old English; see *TOE1*, *TOE2*, *TOE3*, *TOE4*.

TOE1 = A Thesaurus of Old English, eds. J. Roberts et al., 2 vols., King's College London Medieval Studies XI (London, 1995).

TOE2 = A Thesaurus of Old English, eds. J. Roberts et al., 2 vols., Costerus New Series 131, 2nd edn (Amsterdam, 2000).

TOE3 = *A Thesaurus of Old English*, eds. J. Roberts et al. (Glasgow, 2005). http://libra.englang.arts.gla.ac.uk. Accessed: 1 February 2013.

TOE4 = *A Thesaurus of Old English*, eds. J. Roberts et al. (Glasgow, 2015). http://oldenglishthesaurus.arts.gla.ac.uk. Accessed: 15 March 2022.

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.

Wierzbicka, A., *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese*, Oxford Studies in Anthropological Linguistics 8 (New York, 1997).

Appendix 8.A:
EEMEE datasets publicly available

Many of the datasets created in the EEMEE research project have been made available publicly alongside the Linguistic Linked Data version of *A Thesaurus of Old English*. Figure 8.A.1 show the overview of content available in Evoke. Information on the new datasets, including their licenses and availability, is provided below.

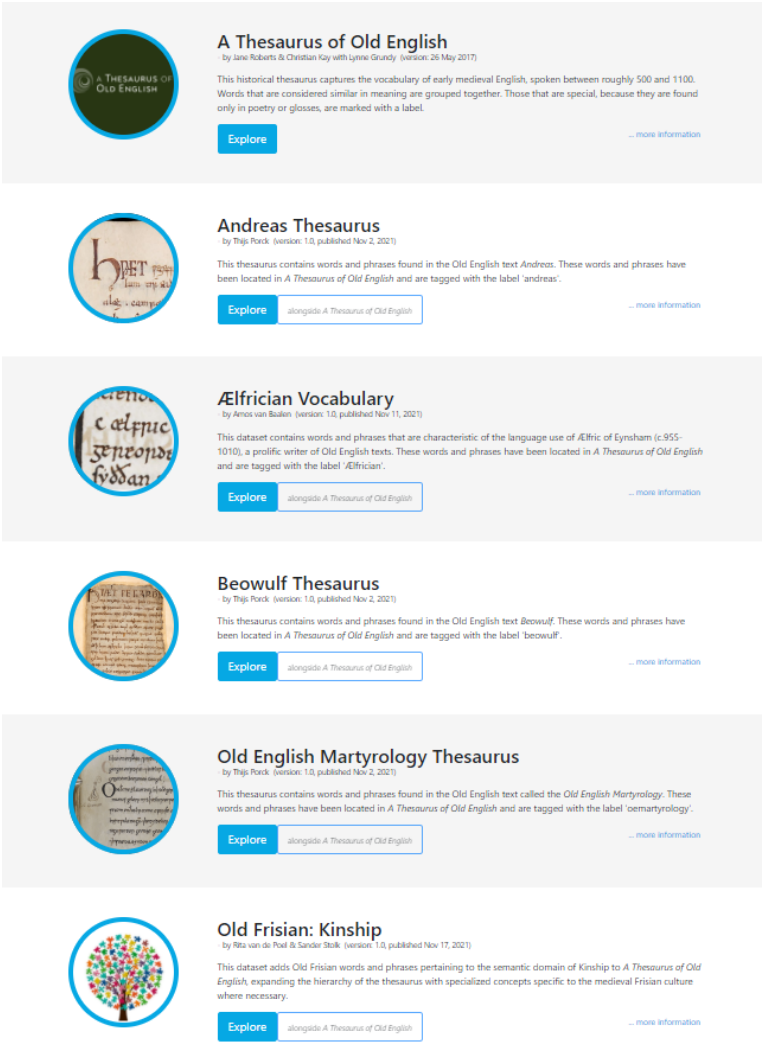


Figure 8.A.1.: Overview of content available in Evoke at http://evoke.ullet.net/content/.

Andreas Thesaurus
Creator: Thijs Porck
Version: 1.0
Released: 2 November 2021
License: CC-BY-SA
Availability:
- in Evoke: http://evoke.ullet.net/content/andreas/
- in DataverseNL: https://doi.org/10.34894/IHVH0Z

Ælfrician Vocabulary
Creator: Amos van Baalen
Version: 1.0
Released: 11 November 2021
License: CC-BY-SA
Availability:
- in Evoke: http://evoke.ullet.net/content/aelfric/
- in DataverseNL: https://doi.org/10.34894/4BQ3ER

Beowulf Thesaurus
Creator: Thijs Porck
Version: 1.0
Released: 2 November 2021
License: CC-BY-SA
Availability:
- in Evoke: http://evoke.ullet.net/content/beowulf/
- in DataverseNL: https://doi.org/10.34894/TOTFGZ

Old English Martyrology Thesaurus
Creator: Thijs Porck
Version: 1.0
Released: 2 November 2021
License: CC-BY-SA
Availability:
- in Evoke: http://evoke.ullet.net/content/oemartyrology/
- in DataverseNL: https://doi.org/10.34894/QZCNW1

Old Frisian: Kinship
Creators: Rita van de Poel and Sander Stolk
Version: 1.0
Released: 17 November 2021
License: CC-BY-SA
Availability:
- in Evoke: http://evoke.ullet.net/content/ofris-kinship/
- in DataverseNL: https://doi.org/10.34894/SOLVNU

Appendix 8.B: Exploring Early Medieval English Eloquence workshops

This appendix details the three workshops within the project 'Exploring Early Medieval English Eloquence' (EEMEE), initially titled 'Exploring Anglo-Saxon Eloquence' (EASE). The project centred around the use of the web application Evoke and *A Thesaurus of Old English* (*TOE*) for use in research and education. Further information on the workshops and other events surrounding EEMEE and Evoke can be found at http://evoke.ullet.net/events. EEMEE was supported by the LUCAS Extra Resources Open Call-II Grant 2020, awarded by the Leiden University Centre for the Arts in Society, and the LUCDH Small Grant 2018, awarded by the Leiden University Centre for Digital Humanities.

8.9.1. First workshop

The first workshop in the series was held at Leiden University on February 1st, 2019. The full programme is shown in Figure 8.B.2. The workshop drew attention to the web application Evoke and explored its value for various fields of research. Researchers convened to formulate potential lines of enquiry that utilise *TOE* in combination with this software. The aim of the first workshop was twofold: (1) to formulate powerful and evocative case studies in research using the thesaurus, and (2) to elicit feature requests to improve the Evoke platform for research purposes. Three talks at the start of the workshop introduced participants to the two resources. The full programmes of the workshops are included on the last three pages of the appendix.

After the introductory talks, the workshop continued with brainstorming on the use of the two resources in research. This activity was facilitated through the Change Pathway from the Europeana Impact Playbook.[58] This framework allowed participants to formulate lines of enquiry in terms of the intended outcomes, any resources required besides *TOE* and Evoke, and the activities involved. Figure 8.B.1 depicts the Change Pathway as filled out for Thijs Porck's proposed research concerning onomasiological profiling of the Old English text *Beowulf*. During the workshop, groups of researchers reflected on each other's ideas, which were sketched in this manner, and suggested improvements. Subsequently, feature requests for Evoke were elicited based on these ideas. On handouts, researchers could describe functionality desired of tooling to benefit research. The results have been incorporated into the development roadmap of Evoke and can be found in Appendix 2.A.

8.9.2. Second workshop

On October 17th, 2020, twelve researchers presented preliminary results of their research utilizing Evoke and *TOE* in the second EEMEE workshop. Its programme is shown in Figure 8.B.3. Old English language and culture, developments of metaphors, lexis used in specific texts or by specific authors,

[58] See 'Europeana Impact Playbook'.

lexicographical practices, ways in which Old Germanic languages can be contrasted with each other, and the use of these digital resources in a classroom setting — the great variation of these talks demonstrated the value of historical language thesauri and their potential in academic research. Discussion of the approaches and results thus far supplied participants with useful notes for refining, and advancing, their work over the next few months.

8.9.3. Workshop at ICEHL-21

A full-day workshop on working *TOE* and Evoke was held at the 21st International Conference on English Historical Linguistics (ICEHL-21) on June 7th, 2021. The papers presented in this workshop demonstrated ways in which these two digital resources, at times complemented with additional data, can be used to explore exciting new aspects of Old English language and culture. Open to all ICEHL-21 attendants, the workshop discussed the results of the EEMEE case studies. The full programme is listed in Figure 8.B.4.

After the workshop, the majority of the case studies were developed into articles and submitted for inclusion in *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4. This special issue of the journal, published in November 2021, is titled 'Exploring Early Medieval English Eloquence: A Digital Humanities Approach with *A Thesaurus of Old English and Evoke*' and was edited by Thijs Porck and Sander Stolk.

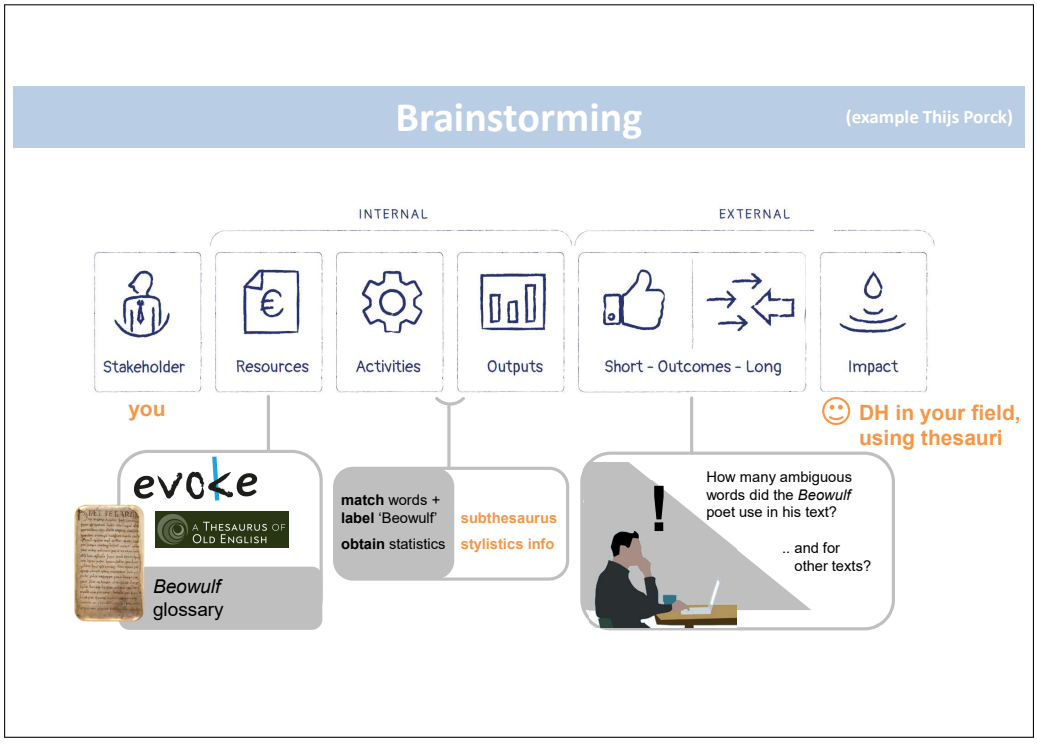
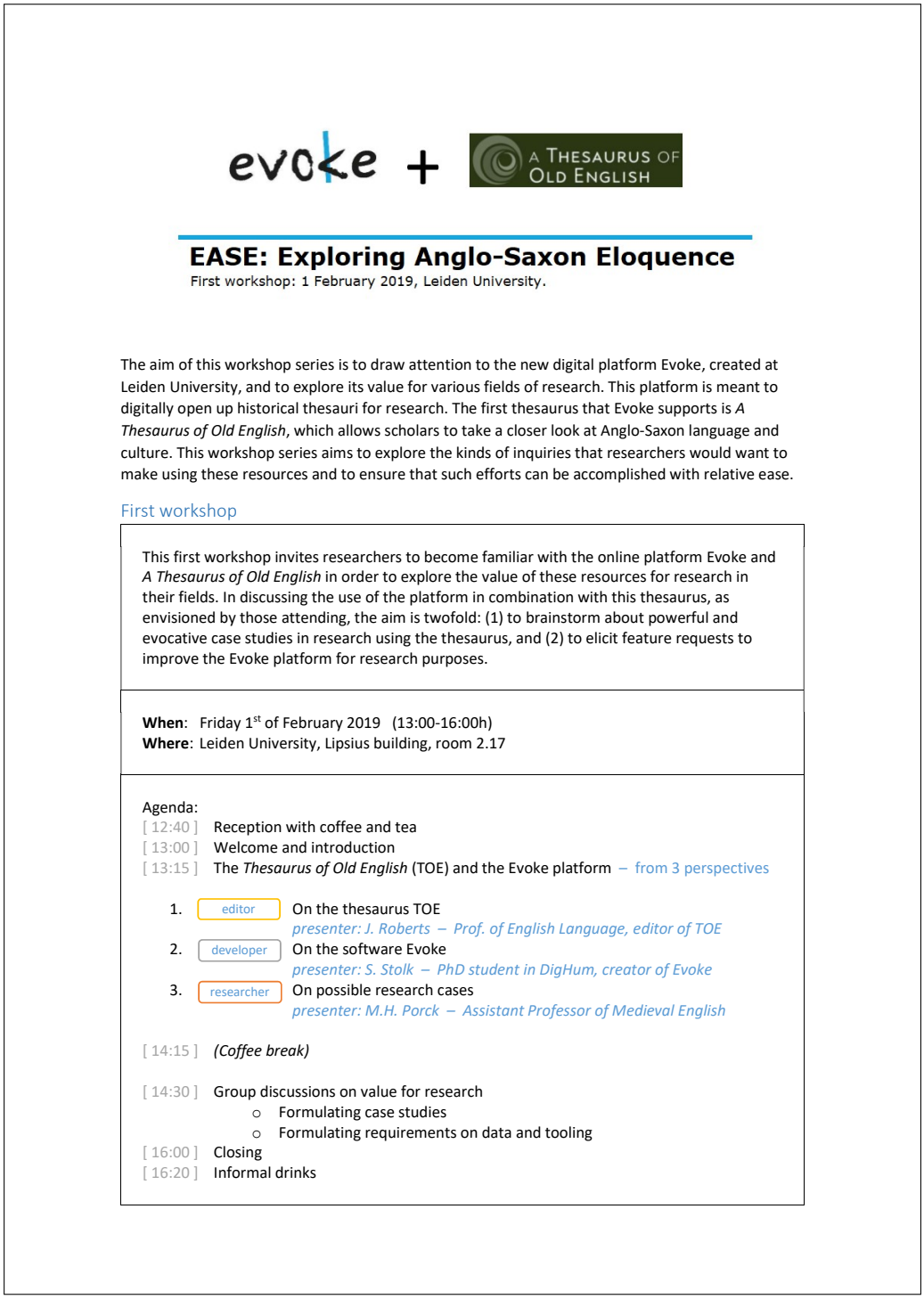


Figure 8.B.1.: Change Pathway towards onomasiological profiling of *Beowulf*.



evoke + A THESAURUS OF OLD ENGLISH

EASE: Exploring Anglo-Saxon Eloquence

First workshop: 1 February 2019, Leiden University.

The aim of this workshop series is to draw attention to the new digital platform Evoke, created at Leiden University, and to explore its value for various fields of research. This platform is meant to digitally open up historical thesauri for research. The first thesaurus that Evoke supports is *A Thesaurus of Old English*, which allows scholars to take a closer look at Anglo-Saxon language and culture. This workshop series aims to explore the kinds of inquiries that researchers would want to make using these resources and to ensure that such efforts can be accomplished with relative ease.

First workshop

This first workshop invites researchers to become familiar with the online platform Evoke and *A Thesaurus of Old English* in order to explore the value of these resources for research in their fields. In discussing the use of the platform in combination with this thesaurus, as envisioned by those attending, the aim is twofold: (1) to brainstorm about powerful and evocative case studies in research using the thesaurus, and (2) to elicit feature requests to improve the Evoke platform for research purposes.

When: Friday 1st of February 2019 (13:00-16:00h)
Where: Leiden University, Lipsius building, room 2.17

Agenda:
[12:40] Reception with coffee and tea
[13:00] Welcome and introduction
[13:15] The *Thesaurus of Old English* (TOE) and the Evoke platform – from 3 perspectives

1. editor — On the thesaurus TOE
   presenter: J. Roberts – Prof. of English Language, editor of TOE
2. developer — On the software Evoke
   presenter: S. Stolk – PhD student in DigHum, creator of Evoke
3. researcher — On possible research cases
   presenter: M.H. Porck – Assistant Professor of Medieval English

[14:15] *(Coffee break)*

[14:30] Group discussions on value for research
- Formulating case studies
- Formulating requirements on data and tooling

[16:00] Closing
[16:20] Informal drinks

Figure 8.B.2.: Programme of the first workshop.

Virtual Workshop 17-10-2020

Working with *A Thesaurus of Old English* and the digital platform *Evoke*

Organisers: Thijs Porck (Leiden University) & Sander Stolk (Leiden University)
Virtual conf.-room: https://smart.newrow.com/#/room/ldt-698 (open in Chrome, Edge or Firefox)

11:00 CEST – Welcome by organisers

11:10 CEST - Session 1: *A Thesaurus of Old English*: Past, present and future

11:10-11:30 – "The pilot study for the Glasgow Historical Thesaurus"
Jane Roberts (King's College London)
11:30-11:50 – "Children and cousins: Projects arising from the Historical Thesaurus"
Marc Alexander + **Fraser Dallachy** (University of Glasgow)
11:50-12:15 – "*Evoke* and *A Thesaurus of Old English*: Exploring language and culture"
Sander Stolk (Leiden University)

12:15- 12:30 – Short break

12:30 CEST - Session 2: Creating your own Old English thesaurus with Evoke

12:30-12:55 – "*By their words you shall know them*: A *Beowulf* thesaurus and onomasiological profiling of Old English texts with *Evoke*"
Thijs Porck (Leiden University)
12:55-13:15 – "Mapping conceptual variation through *A Thesaurus of Old English* and *Evoke*: Towards a topical thesaurus of Old English emotional expressions"
Anas F. Khan (Institute for Computational Linguistics "A. Zampolli", CNR, Pisa) +
Javier E. Díaz-Vera + **Francisco J. Minaya Gómez** (University of Castilla-La Mancha) + **Monica Monachini** (Institute for Computational Linguistics "A. Zampolli", CNR, Pisa)

13:15-14:00 – Lunch break (Bring your own lunch…)

14:00 CEST - Session 3: Evoke and Old English Studies

14:00-14:25 – "*Evoke* and the history of Old English lexicography: Preliminary explorations"
Rachel Fletcher (University of Glasgow)
14:25-14:50 – "*Unhal, unmiht ond wanhal*: Exploring Old English semantic fields of disease and disability"
Berber Bossenbroek (Leiden University)
14:50-15:15 – "*Evoke* and *A Thesaurus of Old English* in the Old English classroom"
Kees Dekker (University of Groningen)

15:15-15:30 – Short break

15:30 CEST – Session 4: Ælfrician English, Old Frisian and Old Dutch

15:30-15:55 – "Using *A Thesaurus of Old English* and *Evoke* to research 'Ælfrician' vocabulary"
Amos van Baalen (Leiden University)
15:55-16:20 – "Contrasting Old Frisian with Old English: An exploration using *A Thesaurus of Old English*, the *Altfriesisches Handwörterbuch*, and the *Evoke* platform"
Rita van de Poel (Leiden University)
16:20-16:55 – "Linking vocabulary of the *Dictionary of Old Dutch* in DiaMaNT to *A Thesaurus of Old English*: An exploration of the possibilities and challenges"
Katrien Depuydt + **Jesse De Does** (Dutch Language Institute)

16:55 CEST – Closing by organisers + possibility for final discussion

Figure 8.B.3.: Programme of the second workshop.

EASE Workshop @ ICEHL-21

Working with *A Thesaurus of Old English* and the digital platform Evoke

organised by **Thijs Porck** + **Sander Stolk** (Leiden University)

Welcome [10:00-10:15 CEST]

Session 1: A Thesaurus of Old English: Past, present and future [10:15-11:45]

"The pilot study for the Glasgow Historical Thesaurus"
Jane Roberts (King's College London)

"Children and cousins: Projects arising from the Historical Thesaurus"
Marc Alexander + **Fraser Dallachy** (University of Glasgow)

"Evoke and *A Thesaurus of Old English*: Exploring language and culture"
Sander Stolk (Leiden University)

Session 2: Creating your own Old English thesaurus with Evoke [12:00-13:00]

"*By their words you shall know them*: A *Beowulf* thesaurus and onomasiological profiling of Old English texts with Evoke"
Thijs Porck (Leiden University)

"Using *A Thesaurus of Old English* and Evoke to research 'Ælfrician' vocabulary in two twelfth-century texts"
Amos van Baalen (Leiden University)

- Lunch -

Session 3: Evoke and Old English Studies [14:30-15:30]

"Mapping conceptual variation through *A Thesaurus of Old English* and Evoke: Towards a topical thesaurus of Old English emotional expressions"
Anas F. Khan (Institute for Computational Linguistics "A. Zampolli", CNR, Pisa) +
Javier E. Díaz-Vera + **Francisco J. Minaya Gómez** (University of Castilla-La Mancha) +
Monica Monachini (Institute for Computational Linguistics "A. Zampolli", CNR, Pisa)

"Evoke and *A Thesaurus of Old English* in the Old English classroom"
Kees Dekker (University of Groningen)

**Session 4: Old English, Old Frisian, Old Dutch –
Comparing Old Germanic lexis with Evoke [15:45-16:45]**

"Contrasting Old Frisian with Old English: An exploration using *A Thesaurus of Old English*, the *Altfriesisches Handwörterbuch*, and the Evoke platform"
Rita van de Poel (Leiden University)

"Linking vocabulary of the *Dictionary of Old Dutch* in DiaMaNT to *A Thesaurus of Old English*: An exploration of the possibilities and challenges"
Katrien Depuydt + **Jesse De Does** (Dutch Language Institute)

Reflection/Closing

Figure 8.B.4.: Programme of the third workshop.

Appendix 8.C: Student-evaluation of Evoke (Nov. 2018)

On 13 November 2018, a small evaluation was carried out on the usefulness of Evoke in education. This evaluation took place in a 2-hour workshop, as part of a third-year Bachelor course on Old English at Leiden University, on digital resources for studying the language and culture. In the workshop, twenty-two students worked on assignments that revolved around the information available in *A Thesaurus of Old English* to study Old English language and culture. The evaluation contrasted the usefulness of Evoke with that of the existent *TOE* website, made available by the University of Glasgow (UoG).[59] The intention was to ensure that the typical classroom setting was disturbed as little as possible (and therefore to avoid the introduction of foreign elements, such as cameras and microphones) whilst still yielding valuable information on the user experience in education.

During the workshop, students were asked to form pairs; one student would work with Evoke, the other with UoG. In the workshop section devoted to *TOE*, each student first got acquainted with the website assigned to them over the time span of five minutes. They were asked to explore how these websites could be navigated and what the options were that they offered, informing each other of their findings afterwards. The subsequent thirty minutes were spent on worksheets with assignments on the *TOE* data. These were to be solved individually, using the assigned website, with the possibility to obtain help from the paired student when needed. After these assignments, both pieces of software were evaluated by the students.

In order to assess the usefulness of Evoke in a classroom setting, fundamental aspects were evaluated that, combined, establish a set of metrics on usefulness. Usefulness can be described as a combination of utility (i.e., the extent to which the application provides the necessary features to support users' needs) and usability (i.e., the ease with which the user interface can be used).[60] Usability, in turn, can be broken down further into various aspects to facilitate a more fine-grained evaluation. Table 8.C.1 lists five key aspects of usability. The classroom setting allowed evaluation of three of the five aspects on usability: learnability, efficiency, and satisfaction. Memorability and errors were not measured, since students would, during this workshop, be getting acquainted with the user interface of Evoke. The three evaluated aspects of usability were included alongside utility and an overall impression on usefulness.

The evaluation of Evoke and UoG, across various aspects of usefulness, was performed through a short poll on each aspect. In order to remove bias from these polls, the evaluation drew on the Microsoft Desirability Toolkit,[61] which consists of a list of 118 words or phrases for possible reactions (e.g., "convenient", "difficult", "boring"). The students were asked to select the words

[59] The evaluation was performed on Evoke v1.2.0 and UoG's *TOE4*.

[60] See, for instance, Nielsen Norman Group's methodologies for evaluating user experience. https://www.nngroup.com/articles/usability-101-introduction-to-usability/

[61] Benedek and Miner, 'Measuring Desirability'.

Aspect	Description
learnability	the ease with which first-time users can accomplish basic tasks
efficiency	the speed with which users can perform tasks
memorability	the extent to which users remember how to work with the interface after not having used it for an extended period
errors	the number and severity of mistakes users make
satisfaction	the pleasantness of the design and its use

Table 8.C.1.: Key aspects of usability

that best described their stance towards the aspect under consideration, which was introduced by a short phrase (e.g., "The look/visuals of that website" for the aspect of satisfaction). Their possible answers, in the form of these reaction words, were narrowed down to a maximum of ten that were suitable for the aspect in question (such as "fast" and "slow" for efficiency).

The results of the evaluation are shown in the tables below for both users of Evoke (ten out the twenty-two students) and UoG (twelve students).[62] A wordcloud next to each table visualizes the results for Evoke, specifically, with scale and darkness of a word or phrase representing the relative number of users that selected it. These results show that both Evoke and UoG were received positively by the students, on both matters of utility and usability. When contrasting the two websites in the results, the most striking differences include that Evoke was more often considered to offer "desirable" functionality, which students later indicated was mostly owing to the statistics generated by the application, and "fun" to use in the assignments.

Word	Value	Evoke users	UoG users
Desirable	+	50 %	17 %
Ineffective	-	10 %	17 %
Powerful	+	20 %	8 %
Helpful	+	100 %	100 %
Dated	-	0 %	0 %
Cutting edge	+	30 %	8 %
Irrelevant	-	0 %	0 %
Not valuable	-	0 %	0 %
Poor quality	-	0 %	0 %
Useful	+	100 %	92 %

Table 8.C.2.: Results for Evoke on utility ("The functionality offered by that website")

[62] The discrepancy with the number of students working with Evoke and with UoG is the result of one pair of students, out of the eleven pairs, having misunderstood their distinctive roles and worked with UoG both.

Word	Value	Evoke users	UoG users
Effortless	+	40 %	17 %
Annoying	-	0 %	25 %
Fast	+	70 %	42 %
Slow	-	10 %	25 %
Disruptive	-	10 %	25 %
Efficient	+	70 %	50 %

Table 8.C.3.: Results for Evoke on efficiency
("The efficiency with which it allowed me to perform the tasks")

Word	Value	Evoke users	UoG users
Difficult	-	0 %	0 %
Straightforward	+	60 %	50 %
Confusing	-	20 %	33 %
Too technical	-	0 %	0 %
Clear	+	80 %	42 %
Incomprehensible	-	0 %	0 %
Accessible	+	60 %	58 %
Understandable	+	50 %	75 %
Easy	+	70 %	33 %
Stressful	-	0 %	8 %

Table 8.C.4.: Results for Evoke on learnability
("The process of learning to use that site")

Word	Value	Evoke users	UoG users
Attractive	+	50 %	25 %
Boring	-	10 %	0 %
Clean	+	90 %	58 %
Overwhelming	-	0 %	0 %
Calm	+	60 %	58 %
New	+	20 %	0 %
Cutting edge	+	0 %	0 %
Unattractive	-	0 %	0 %
Patronizing	-	20 %	0 %
Old	-	0 %	8 %
Organized	+	80 %	92 %
Satisfying	+	50 %	25 %

Table 8.C.5.: Results for Evoke on satisfaction
("The look/visuals of that website")

Word	Value	Evoke users	UoG users
Convenient	+	70 %	67 %
Frustrating	-	10 %	33 %
Valuable	+	20 %	25 %
Useful	+	80 %	100 %
Poor quality	-	0 %	0 %
Essential	+	20 %	8 %
High quality	+	30 %	8 %
Dated	-	0 %	0 %
Fun	+	90 %	25 %
Professional	+	20 %	50 %

Table 8.C.6.: Results for Evoke on overall perception
("My feeling of that website overall...")

Chapter 9

Information:
The article that starts on the next page was published in 2021, available open access under the CC BY license. The only change to that paper, here, is the inclusion of two numbers for the benefit of readers of the dissertation: the overall page number and the chapter number (presented in the margin in a grey box and a white box, respectively). When citing, please refer to the original publication and its page numbering.

Publication:
Rita van de Poel and Sander Stolk, 'A Case of Kinship: Onomasiological Explorations of KINSHIP in Old Frisian and Old English'. *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021), pp. 457-492. doi: 10.1163/18756719-12340239.

BRILL

AMSTERDAMER BEITRÄGE ZUR
ÄLTEREN GERMANISTIK 81 (2021) 457–492

ABäG
brill.com/abag

A Case of Kinship

Onomasiological Explorations of KINSHIP in Old Frisian and Old English

Rita van de Poel | ORCID: 0000-0001-5960-7913
Leiden University Centre for the Arts in Society, Leiden University, Leiden, Niederlande
g.i.van.de.poel@hum.leidenuniv.nl

Sander Stolk | ORCID: 0000-0003-2254-6613
Leiden University Centre for the Arts in Society, Leiden University, Leiden, Niederlande
s.s.stolk@hum.leidenuniv.nl

Abstract

This article describes onomasiological explorations of Old Frisian and Old English lexis in the semantic field of KINSHIP through a novel, digital approach. In connecting Old Frisian lexis, drawn from the *Altfriesisches Handwörterbuch* (*AFWB*), to the overarching structure of *A Thesaurus of Old English* (*TOE*), a dataset has been created that shares a semantic framework with the one existing for Old English lexis. The connected resources are shared and analysed using the web application Evoke. Statistical data provided by this tool, such as the degree of lexicalization for this field, facilitates comparative analyses of the two historical languages. As this article demonstrates, the reuse of the onomasiological macrostructure of *TOE* offers new insights into linguistic and cultural aspects of these two languages and their language communities.

Keywords

onomasiology – KINSHIP – Old Frisian – Old English – Evoke – digital humanities – linguistic linked data – thesaurus

© RITA VAN DE POEL AND SANDER STOLK, 2021 | DOI:10.1163/18756719-12340239
This is an open access article distributed under the terms of the CC BY 4.0 license.

1 Introduction

Since its publication in 1995, *A Thesaurus of Old English* (*TOE*) has been an asset to research into Old English language and culture. This lexicographic resource captures the early medieval English lexis (c.500–1100), which it does not order alphabetically, like most lexicographic works for Old English do, but onomasiologically: words are organized by means of an overarching topical structure, allowing users to go from meaning to words that express that meaning. In thus positioning words that are similar in meaning close to one another, including the grouping of synonyms, *TOE* has facilitated numerous word studies and semantic field studies[1] and its contribution to Old English studies has been met with high praise from scholars (Dance, 1997: 312; Görlach, 1998: 398–399). Hence, the value of an onomasiological ordering of lexis has been demonstrated for Old English. As can be imagined, that value is not limited to this specific language.

Comprehensive thesauri such as *TOE*, in spite of their value for research, are scarce. The lack of such resources for entire languages is unsurprising: creating a thesaurus takes a considerable amount of time and effort. To illustrate, the creation of *TOE* (discussed in the contribution by Jane Roberts in this special issue) has taken a team of researchers, themselves drawing from available dictionaries of Old English, over fifteen years (Roberts, 1978). Amongst the languages which have hitherto not yet been captured in a thesaurus is Old Frisian. Yet, in this particular case, the characteristics of this language, combined with resources currently available, may pave the way towards an onomasiological ordering of its lexis in a shorter time than was needed for the creation of a thesaurus of Old English.

Parallels between Old English and Old Frisian have often been drawn and tend to be mentioned explicitly in books that serve as introductions to these languages (see Bremmer, 2009: 125–128; Baker, 2012: 7–8). Indeed, similarities between these 'sibling languages' exist on multiple levels – including their lexis. These parallels suggest that an onomasiological marcrostructure designed for Old English may form a good starting point for Old Frisian. Linking Old Frisian lexis to *TOE*, positioning words and word senses in appropriate locations of its macrostructure, may yield an onomasiological perspective on Old Frisian that is currently absent. Moreover, the placement of Old Frisian senses in

1 E.g. Hough (2007); Díaz-Vera (2011); Rauer (2017); Tissari (2017); Ojanguren López (2021); Scott (2021).

this thesaurus is likely to also enable a comparison between these two closely related languages on onomasiological and lexico-semantic levels.

In this article, we describe preliminary work in connecting the Old Frisian lexis to *TOE* and the use of the results to compare Old Frisian with Old English. For this purpose, we have used the Old Frisian lexis from one particular semantic field, namely that of KINSHIP, mainly represented in *TOE* by the category "02.03.02 Family/household".[2] The connected resources are shared and analysed using the web application Evoke (Stolk, 2018). This application allows researchers to browse and analyse *TOE* alongside additional content. In addition to discussing our findings in linking up Old Frisian information to the thesaurus, we will present comparative analyses of Old Frisian and Old English provided by the statistical data that Evoke offers, such as the degree of lexicalization of this specific field. Thus, we hope to answer, and nuance, the following questions regarding the semantic fields of KINSHIP in Old Frisian and Old English:

1. Can the Old Frisian lexis be allocated to the onomasiological macrostructure of *TOE*?
2. Can Evoke, in combination with *TOE* and the linked Old Frisian lexis, offer new insights into linguistic and cultural aspects of Old Frisian and, in contrast with Old Frisian, Old English?

2 Related Work

The number of studies pertaining to Old Frisian linguistics is significantly smaller than those on Old/Middle English or Old/Middle High German (Bremmer, 2007: 55). Most Old Frisian linguistic research investigates a relatively limited aspect of the language or the texts: extensive, comprehensive research is scarce (Bremmer, 1992; 2021). Lexico-semantic Old Frisian studies are almost without exception short treatises on the characteristics of specific lexemes; only a few publications take an onomasiological perspective or comprise an entire semantic field.[3] Although these studies in Old Frisian provide

2 "Family/household" in *TOE*. Identifier: 1108. Location: 02.03.02. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1108.

3 Noteworthy studies are Munske's seminal work on the Germanic vocabulary in the field of crime, in which he notes the parallels between the terminology in Old East Frisian and Old English (1973), Bremmer's work on verbal injury in late medieval Frisia (1998), Popkema's on the conceptualization of oaths (2007), and Hofmann's monograph on medieval everyday life as conceptualized in Old Frisian (2015).

valuable insights into lexical expressions and their meanings, the lexico-semantic research in Old Frisian has hitherto mostly taken a traditional philological approach: they lack empirical or quantitative components.[4]

Research pertaining Old Frisian language is not only sporadic but also often undervalued – especially in the field of historical linguistics Old Frisian is considered to be neglected (Bremmer, 2009: 18; Stroh, 1985: 371; Munske, 2001: xiii; Salmons, 2007: 367). Digital environments for studying Old Frisian are scarce, hampering innovative research (such as comparative studies) of Old Frisian.[5] Recently, however, a digital, lemmatized, and representative corpus of Old Frisian has been published online, a welcome lexicographic resource that can be used by both researchers and students (Van de Poel, 2019). This *Corpus Oudfries* contains a large sample of the Old Frisian language and can be searched on three linguistic levels (tokens, lemmata, and parts of speech) and on extra-linguistics levels (i.e. dialect and date). At this point in time, the corpus contains 235,462 tokens and 177 text witnesses from 11 manuscripts. Digital resources, such as this corpus and Evoke, provide new means for research into Old Frisian language and culture.

A semantic study of the domain of KINSHIP in Old Frisian has, to our knowledge, not been carried out before, although a limited number of studies are available that investigate individual lexemes[6] and/or certain medieval Frisian legal aspects (Meijers, 1925; Kok, 1947; Boersma, 1961; Sterringa, 1998). For Old English, a comprehensive lexical investigation of kinship terminology has not been conducted either. Various studies have, however, been realized by anthropologists, traditionally the scholars that contributed the most to work on kinship systems and terminology. Lorraine Lancaster (1958), Henry Loyn (1974), and Georg Pfeffer (1987) are among the most influential scholars to have explored Anglo-Saxon kinship terms. Furthermore, certain lexical elements or subdomains within KINSHIP have been investigated for Old English (Spolsky, 1977; Bremmer, 1980; Lowe, 1993; Bajema, 1994; Fischer, 2006; Durkin, 2019).

Relevant semantic word field studies on KINSHIP that should be mentioned are descriptive and comparative studies by Ariane Diepeveen (2003)

4 An exception is Versloot's dissertation, which incorporates quantitative analyses to chart language change in 15th century West Frisian (2008).

5 "The creation of a lemmatized (…) Old Frisian corpus would be invaluable for this investigation, as for so many others" (Colleran, 2017: 102). "Many instruments historical linguists have been accustomed to whilst studying older language stages are lacking or poorly available for Old Frisian" (Boutkan and Siebinga, 2005: vii).

6 E.g., Meijering (1985) discusses the etymology of *kind*, Bammesberger (1968) looks into the etymology of *swager*, and Pospelova (2018) examines Indo-European rules for compounding in Old Frisian kinship terms.

and Susanne Zeilfelder (2015). Diepeveen provides a diachronic outline how the kinship terms of the (North and West) Germanic languages have evolved over the centuries. She includes almost all Germanic languages in her work, including Old English, but regrettably Old Frisian is absent. Zeilfelder's onomasiological work mainly focuses on the semantic development and etymology of German. She provides comparisons with and context from other Indo-European languages and to this end also describes Old English and Old Frisian lexemes. However, she does not do this consistently for each sense and/or cognate word. The current study intends to work towards filling this lacuna and to explore how an onomasiological approach, and tools such as Evoke, can facilitate studies of Old Frisian terminology for KINSHIP.

3 Background: Old Frisian

Old Frisian is the term used to indicate the earliest version of the Frisian language, written between c.1100–1550. Only a fairly limited number of Old Frisian manuscripts (18) have come down to us, of which the texts are almost entirely juridical in nature.[7] Besides the major juridical text collections, the largest body of Old Frisian texts consists of administrative documents (charters, wills, deeds), which represents about 50% of the entire collection of written Old Frisian words.[8]

The denominator "Old" for Old Frisian suggests that it was more archaic than its contemporary neighbouring languages (such as Middle English and Middle High German) and that it was linguistically compatible with, for example, Old English, Old High German, or Old Saxon. Scholars of Old Frisian have investigated the periodization of Old Frisian, which is complicated by the fact that dialectological differences in the language of the textual witnesses concur with the chronological differences: "some linguistic differences have to be interpreted in a chronological sense and some in a dialectological sense" (Versloot, 2004: 256). Rolf Bremmer summarizes the various chronological and dialectological research on Old Frisian and indicates that "the way in which languages are divided into periods depends on the criteria which one chooses in order to establish such periods" (2009: 125).

7 Bremmer (1992: 6–15) provides an overview of the most important Old Frisian texts and manuscripts.

8 According to Vries (2001: 594) 40%, but closer to 50% based on token number estimates (own research, unpublished).

It has generally been accepted that medieval Frisian texts can be divided into an older and a younger variety (classical/old vs. post classical/late)[9] and that these periods more or less overlap with the division into regions or dialects. Arjen Versloot indicates that the dialect variety East Old Frisian dates mainly to the period 1300–1450 and was written in the present Dutch province of Groningen and in the German region of East Friesland (2004: 285). The second group originates from the province of Fryslân and was written in the period 1450–1525. He concludes: "Whatever periodisation scheme one prefers, the central conclusion is that the oldest Frisian attestations in the manuscripts represent a language that is fairly compatible with other Old Germanic languages" (2004: 285).

4 Methodology

In this section we discuss our methodology for linking the Old Frisian lexis to *TOE* as available in Evoke. In another article in this special issue, Katrien Depuydt and Jesse de Does describe their approach to a similar goal. They have experimented with a (semi-)automated method that matches Old Dutch words from the semantic field of KINSHIP with *TOE*. Our method, in contrast, involves manually matching dictionary senses to thesaurus categories and can be divided into the following steps:

1. Identifying Old Frisian lemmata on KINSHIP in the used source dictionary for Old Frisian.
2. Sense alignment: analysing the different senses of each lemma and matching these senses with a category in *TOE* or introducing a new category, resulting in a semantic classification.
3. Processing the alignment into Linguistic Linked Data and importing the work into Evoke.

4.1 *Identifying Old Frisian Lemmata on KINSHIP*

Creating a list of lemmata belonging to the semantic field of KINSHIP was accomplished by manually searching the *Altfriesisches Handwörterbuch* (*AFWB*) and marking the words that belong to the semantic field of KINSHIP. The concise dictionary *AFWB* covers the Old Frisian language from 1200–1550 and was compiled between 1959 and 2008 by Dietrich Hofmann and Anne Popkema. The Old Frisian words are provided with translations into Modern German. The compilers of *AFWB* did not intend the dictionary to be comprehensive,

9 The terms "classical" and "post-classical" were coined by Sjölin (1966), who also raised the issue of the overlap between dialects and chronology.

but included as many lemmata and sense distinctions as needed for its use as a reading companion to Old Frisian texts (*AFWB*: xxiii–xxiv). The dictionary contains 11,254 headwords, of which 247 (2%) were selected as they belonged to the semantic field of KINSHIP and were therefore eligible for linking to the *TOE*.

A dictionary entry in the *AFWB* has the following format: the lemmata are printed in bold, a reference to dialect is printed in superscript (WL or OL), followed by the part of speech category, senses, sources (there is no indication which sense was found in which text source), composites and cross-references.

> **frouwe, frowe, frou**$^{\text{WL}}$ f. 1) Frau; 2) Ehefrau; 3) Edelfrau, Fürstin; 4) Herrin; 5) die Jungfrau Maria; 6) Schwiegermutter –
> Bas, BBr-D, BDg-U, BEm-E1E2, BFi-F, Bgr-J, BHm-J, BHua-H, BHub-H, (…).
> Komp.: ethelinges-, gā-, hāved-, hērskipes-, hūs-, jest-, jung-, klāster-, londes-, lond-, munekfrouwe

The head word is followed by subsidiary lemmata that pertain to phonological and/or orthographical (and often dialectal) variations of the head word. The lemmata within an entry are arranged in chronological order, which means that the earliest attestation is always the head word. In our methodology, the first lemma was selected for inclusion in the alignment. Thus, the earliest Old Frisian word forms have been imported into Evoke. The compilers of the *AFWB* have harmonized the spellings of many word forms to ensure that lexemes belonging together can easily be retrieved.

An *AFWB* lemma can have multiple senses, which are distinguished by numbers. Frequently, however, the senses of a lemma also contain commas that seem to indicate slightly overlapping meanings. Examples are: *efterswesterling* – "Andergeschwisterkind, Verwandter im dritten Grade" [second cousin, third-degree relative] and *niftakind* – "Grosskind, kind des Nichten" [grandchild, niece's child]. For the scope of the present study, we decided to consider these descriptive meanings as elaborations rather than distinct senses.

4.2 *Sense Alignment*

Three spreadsheets were developed in order to facilitate data entry and subsequent data conversion into Linguistic Linked Data, the interoperable data format used by Evoke. These spreadsheets represent the three main elements that are to be captured: lexical entries, senses, and concepts. Figure 1 shows the sheet for lexical senses. This sheet, which is pivotal in the alignment of the Old Frisian words and their senses with *TOE*, provides local identifiers for the current word (B) and for this word in a specific sense (A), the head word (D), the language code according to ISO 639 (E; here "ofs" for Old Frisian), and the

	A	B	C	D	E	F
1	id	entry-id	[link]	word	language	concept-id
2	aftberen-s1	aftberen	link	aftberen	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1120
3	aldaêm-s1	aldaêm	link	aldaêm	ofs	14
4	aldafeder-s1	aldafeder	link	aldafeder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1183
5	aldemôder-s1	aldemôder	link	aldemôder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1184
6	alderlâs-s1	alderlâs	link	alderlâs	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1124
7	alder-s3	alder	link	alder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1183
8	alder-s4	alder	link	alder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1185
9	alder-s5	alder	link	alder	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1109
10	âthum-s1	âthum	link	âthum	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1262
11	âthum-s2	âthum	link	âthum	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1262
12	basterdbrôther-s1	basterdbrôther	link	basterdbrôther	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1143
13	basterd-s1	basterd	link	basterd	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1141
14	basterdswester-s1	basterdswester	link	basterdswester	ofs	3
15	bênene burch-s1	bênene burch	link	bênene burch	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1787
16	bêneteburch-s1	bêneteburch	link	bêneteburch	ofs	http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1787

FIGURE 1 Spreadsheet used in recording Old Frisian senses and the concepts to which they relate

identifier for the semantic category (F) in which the sense is to be positioned. The identifier of a category can be either an existing one from *TOE* (a web address also referred to as an IRI) or a newly coined category identified by a number and defined on the sheet for lexical concepts – including where this newly defined category fits into the *TOE* taxonomy.

The following activities are involved in assigning the senses of the Old Frisian lexemes to appropriate categories in *TOE* (either existing ones or new ones that add a further degree of specialization):

- Record the lemma and its Modern German senses in the spreadsheets.
- Translate Modern German sense definitions into Modern English.[10]
- Locate suitable *TOE* categories by
  a) browsing the taxonomy of *TOE*
  b) searching for categories that contain keywords from the translated Modern English definitions of the lemma, and
  c) searching for the Old English cognates, if any, and marking the *TOE* categories at which they are positioned.
- Record matching *TOE* categories in the spreadsheets. When no matching *TOE* category is available for a sense, create a new category in the spreadsheet and position that category in the *TOE* taxonomy by recording its superordinate category.
- Determine whether Old English cognates appear in more than one category, since this could imply that the Old Frisian lexeme under investigation would also have to be assigned to these other categories in order to facilitate contrasting the two languages.[11]

10 *DWDS* has been used to examine the German translations.

11 We elaborate on choices made during the classification process in Appendix A.

4.3 *Processing the Alignment for Use in Evoke*

In order to transform the three spreadsheets to Linguistic Linked Data, we have employed the conversion tool OpenRefine along with its RDF plugin. The conversion logic for these sheets has been made publicly available.[12] Each row in the sheet for lexical senses is transformed into an instance of a data element as defined in OntoLex, an interoperable model that has been designed specifically for capturing linguistic data, such as lexical entries and their senses.[13] The resulting Linguistic Linked Dataset has been imported into the online repository of Evoke.

269

9

5 Results

Numbers on the created dataset, which is now publicly available in Evoke, are as follows: 280 lexical senses on KINSHIP, from 247 Old Frisian lemmata, have been aligned with *TOE* categories (see Appendix B).[14] The majority of these senses have been allocated to the semantic field "02.03.02 Family/household", as the following overview shows:

02.01 Existence, life (id: 661)	21 senses
02.03 Humankind (id: 1059)	2 senses + those in subfields (below)
02.03.01 People (id: 1065)	31 senses
02.03.02 Family/household (id. 1108)	215 senses
12.09 Marriage, state of marriage (id: 18602)	11 senses

Since "02.03.02 Family/household" represents the core of Old Frisian terminology on KINSHIP, this case study concentrates its onomasiological analyses on this semantic field.

Old Frisian senses placed under "02.03.02 Family/household" have either been allocated to already existing *TOE* categories (132 senses to 70 *TOE* categories) or to categories newly introduced into the *TOE* taxonomy (83 senses to 57 new categories). Originally, the field "02.03.02 Family/household" contained a total of 175 categories with 324 recorded Old English senses. An overview

12 https://github.com/ssstolk/oldfrisian-kinship/.

13 OntoLex-Lemon: Lexicon Model for Ontologies, *W3C*. http://www.w3.org/2016/05/ontolex/ (10 May 2016).

14 The dataset "Old Frisian: Kinship", containing 247 Old Frisian lemmata, is now publicly available in Evoke: http://evoke.ullet.net/content.

TABLE 1 Item counts within the field "02.03.02 Family/household"

	Old English	Old Frisian
number of lemmata	294	200
number of senses	324	215
number of categories w/ senses allocated to them	175	127

of these numbers is provided in Table 1. A substantial number of categories from this field in the expanded taxonomy have solely Old Frisian or solely Old English senses assigned to them (162 out of a total of 232 categories, or 70%). In the field of KINSHIP, then, the recorded vocabularies of these kindred languages contain many differences in denotations and nuances of words. We will elaborate on some of the more apparent differences in our discussion of the distribution of Old Frisian lexis over the various semantic subfields of "02.03.02 Family/household", in section 6.4.

5.1 *Locating Old Frisian Words, Their Synonyms, and Cognates*

As a consequence of the categorization of the Old Frisian lexis with the macrostructure of *TOE*, the web application Evoke can offer scholars a seamless integration of Old English and Old Frisian lexis for the field of KINSHIP. Thus, not only words for a given concept can be obtained for either language, but also synonyms and possible translations between them. Such an integrated overview of this information can be activated by selecting both relevant datasets in Evoke (i.e. *TOE* and the Old Frisian dataset newly created for this research). When subsequently opening a category such as "02.03.02.03.03 Forefather, ancestor" in the user interface of the application,[15] it is revealed which words were used to express this concept in both Old Frisian and Old English. The list presented in Figure 2 shows six different Old English words for this concept (including *ēr̄fæder* and *ieldra*) compared to three for Old Frisian (viz. *alder*, *forefeder*, *forefirdera*). These words are grouped by language and sorted alphabetically.

The integration of Old Frisian and Old English into Evoke facilitates the comparison of the relationship between the lexicons of both languages. Cognates are words within the same language or in different languages that have a

15 "Forefather, ancestor" in *TOE*. Identifier: 1178. Location: 02.03.02.03.03. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1178.

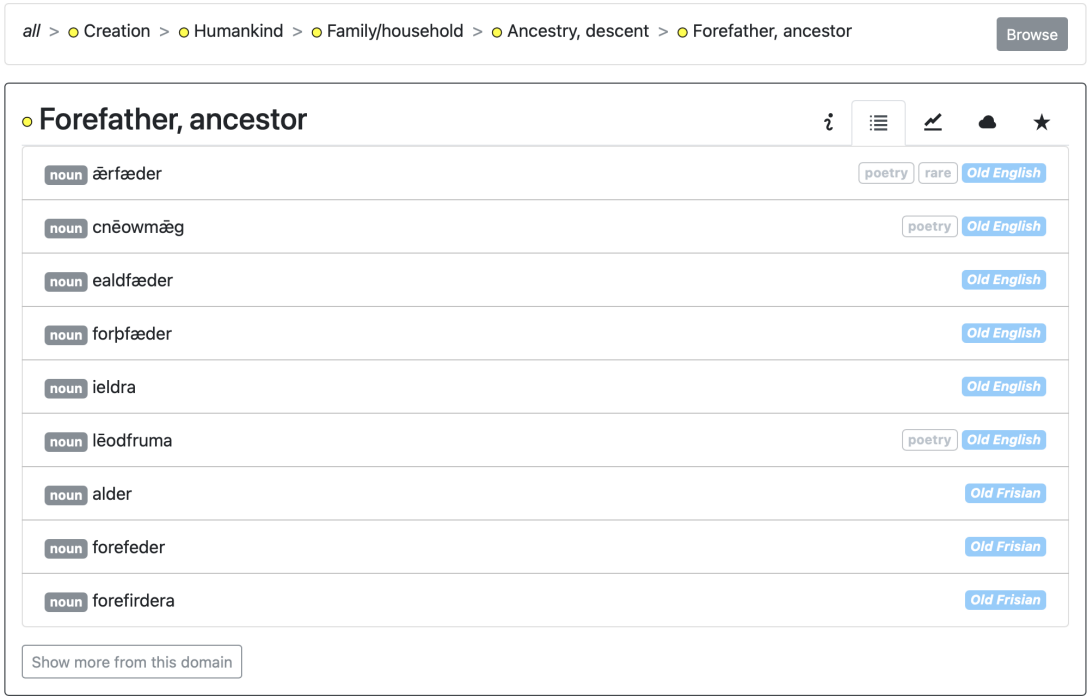


FIGURE 2 List in Evoke of Old English and Old Frisian words denoting "02.03.02.03.03 Forefather, ancestor"

common etymological origin, and therefore resemble each other to a greater or lesser extent in form (Schmitt, 1997: 209; Otwinowska-Kasztelanic, 2011: 4). Awareness of cognates enhances the ability to learn another language – in this case, learning Old Frisian will be easier for someone who is familiar with Old English, and vice versa (Schmitt, 1997: 209; Otwinowska-Kasztelanic, 2011: 4–5).[16] Additionally, finding cognate words in a set of languages is the first step in the comparative method for historical linguists, allowing them to study the development of languages and the reconstruction of common ancestors (Baldi, 2011: 1–16; Trask, 2015: 198–233). Figure 3 lists the various synonyms (in Old English as well as Old Frisian) for Old English *ealda fæder*. Here, Old Frisian *aldafeder* is a cognate of the Old English word that is closest in form: *ealda fæder*. Similarities such as these, i.e. in both form and meaning, facilitate detection of cognates.

5.2 *Kinship Terminology: Cultural Lexical Research of Cognates*

Onomasiological ordering of lexis can be useful for cultural lexical research. Kinship terms are "ways in which people classify their kinship universe" and as such provide clues to the nature of a kinship system in a society as well as

16 Summary of other cognates research in Friel and Kennison (2001).

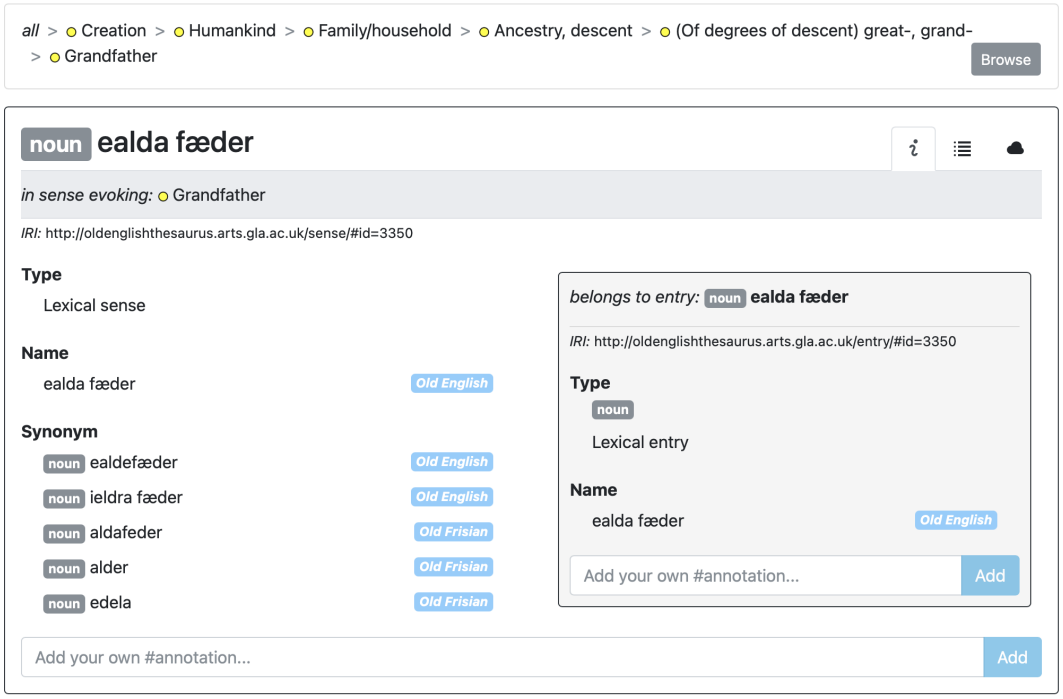


FIGURE 3 Information in Evoke on Old English *ealda fæder* in the sense of "02.03.02.03.04 | 01 Grandfather"

to the social statuses and roles of kinsmen (Fox, 1984: 243). Similar cultures often have very similar reference terms for relatives. It would go beyond the scope of this article to perform an entire analysis of the semantic field in question. However, to illustrate the usefulness of Evoke in comparing Old Frisian and Old English we undertake an exploratory comparative study of cosanguineal KINSHIP terms. We have taken inspiration from well-known research by Lancaster on kinship terminology (1958). Her kinship tree graph, which contains cosanguineal nomenclature in Old English, has been expanded here with corresponding Old Frisian lexis (see Figure 4). The graph, using a genealogical structure, contains nodes and lines to indicate individuals and relations of descent, respectively.[17] For every node in the graph, Evoke has been employed to locate the corresponding Old English and Old Frisian words. The results are shown in Table 2.

17 It should be noted that genealogical structure of this graph differs from the structure found in *TOE*, which is onomasiological in nature and contains more than consanguineal terms.

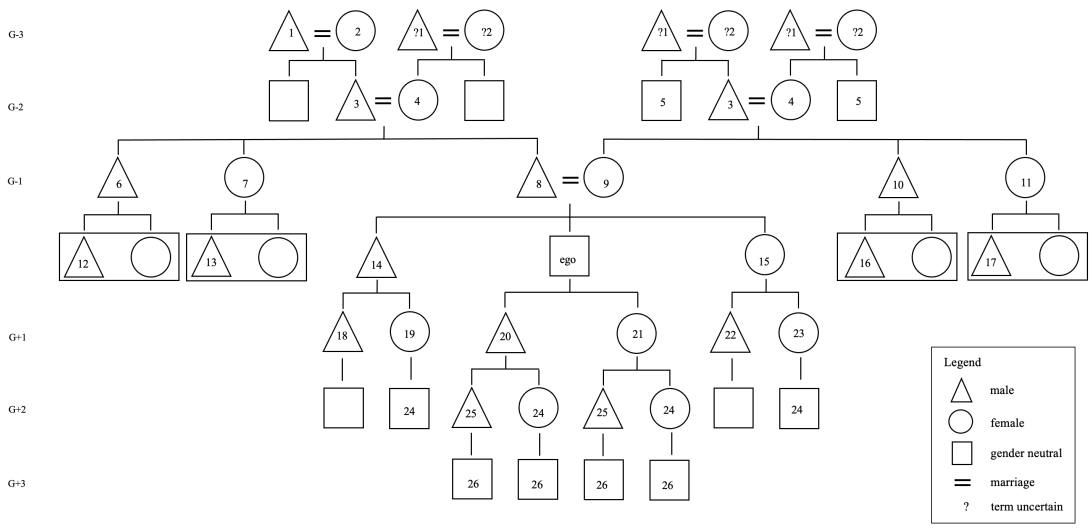


FIGURE 4 Kinship relations

TABLE 2 Cosanguineal kinship terms in Old English and Old Frisian

No.	Old English	Old Frisian	Modern English
1	*ieldra fæder, thridda fæder*	*edela, üraldafeder, alder*	great-grandfather
2	*thridde mōdor*	*eldramōder, üraldemōder*	great-grandmother
3	***ealde mōdor***	***aldemōder***	grandmother
4	***ealdafæder***, *ieldra fæder, ealda fæder*	***aldafeder***, *edela, alder*	grandfather
5		*aldaēm*	granduncle
6	***fædera***	***federia***, *federesbrōther*	uncle, father's brother
7	***fathu***	***fethu***, *federswester*	aunt, father's sister
8	***fæder***	***feder***	father
9	***mōdor***, *ācennicge, bearncennicge, cennestre, byrthe*	***mōder***	mother
10	***ēam***	***ēm***, *mōderesbrōther*	uncle, mother's brother
11	***mōdri(g)e***	***mōdire***, *mōie*	aunt, mother's sister
12	*mōdri(ge), (ge)swēor, geswigra*	*federiasune*	male cousin (father's brother's son)
13	***fathusunu*** *mōdri(ge), (ge)swēor, geswigra*	***fethansune***	male cousin (father's sister's son)
14	***brōthor***	***brōther***	brother
15	***sweostor***	***swester***	sister

TABLE 2 Cosanguineal kinship terms in Old English and Old Frisian (*cont.*)

No.	Old English	Old Frisian	Modern English
16		*emka, emessune*	male cousin (mother's brother's son)
17		*mōdiransune*	male cousin (mother's sister's son)
18	(*ge*)***nefa, brōthorsunu,*** *suhterga*	***neva, brōthersune,*** *brōtherbern, brōtherskind, swesternabern*	nephew
19	*nefene,* ***nift,*** ***brōthordohtor***	***nifte, nifke, brōtheresdochter,*** *brōtherbern, brōtherskind, swesternabern*	niece
20	***sunu, bearn,*** *byrdling, byre, tūdor, eafora, geēacnung*	***sune***, ***bern***, *kind*	son, child (general term)
21	***dohtor***, ***bearn***, *byrdling, byre, tūdor, eafora, geēacnung*	***dochter***, ***bern***, *kind*	daughter, child (general term)
22	(*ge*)***nefa, sweostersunu***, ***sweosterbearn***	***neva***, ***swestersune***, ***swester(na)bern***, *swesterkind, swesterling*	nephew
23	*nefene,* ***nift***, ***sweosterbearn***	***nifte***, ***nifke***, *swesterdochter,* ***swester(na)bern***, *swesterkind, swesterling*	niece
24		*niftlīn, niftakind*	niece's child
25	grandson: *sunsunu, nefa* granddaughter: *nefe, nift*	*bernesbern, kindeskind*	grandchild
26	great granddaughter: *thridde dohtor* great grandson: *thridda sunu*	*kindeskindeskind*	great-grandchild

Comparison of the KINSHIP terminology clearly demonstrates the close relationship between Old English and Old Frisian: cognate forms for similar terms in Table 2 appear in boldface. Old English and Old Frisian have cognates for the lexis for: father (Fa), mother (Mo), brother (Br), sister (Si), son (So), daughter (Da), child, grandfather, grandmother, maternal uncle (MoBr) and aunt

(MoSi), paternal uncle (FaBr) and aunt (FaSi), nephew and niece. Terms for some other blood relations likewise show similar cognate (compound) forms, i.e. great grandfather, greatgrandmother, cousins.

Old Frisian possessed terms for kinship relations that are not found in Old English: *mōdiransune, emessune, aldaēm*. When no Old English lexeme is recorded for a specific sense, it should not be inferred that the concept as such was absent in Old English. Notions such as "father's brother's son" and "mother's sister's son" exist in Old English, but are not lexicalized. Instead, they were expressed with genitival phrases (*fæderan sunu* and *modiran sunu*).

275

9

6 Analysis

Based on the data from *TOE* and the newly created dataset, this section presents a detailed analysis of both the Old English and the Old Frisian lexis located under the semantic field of "02.03.02 Family/household" through the use of the web application Evoke. Evoke offers quantitative information from *TOE*, possibly in combination with additional datasets, in two forms: (1) basic statistics for a specific category and (2) advanced statistics that incorporate the onomasiological structure of *TOE* more fully, which also allow for queries to be customized.

6.1 *Analysis of Parts of Speech Distribution*

The basic statistics of Evoke allow us to provide some insight into matters such as the distribution of the parts of speech within the semantic field of KINSHIP, represented by the *TOE* category "02.03.02 Family/household" and all its subordinate categories. Figure 5 shows the distributions for Old English senses and of Old Frisian ones. When contrasting these numbers, the percentages of nouns for Old English and Old Frisian turn out to be comparable. However, Old Frisian has relatively fewer adjectives and more verbs, adverbs, and phrases than Old English. The marked difference between the relative number of verbs and that of adjectives is especially striking and merits further research.

6.2 *Degree of Polysemy*

The advanced statistics section of Evoke renders, amongst others, a graph that indicates polysemy: the number of senses attributed to a lemma. Indeed, polysemy (and homonymy) can be a measure of the ambiguity of words, demanding the interpreting party to reflect carefully on the intended meaning in an utterance (Chandler and Munday, 2016: s.v. *polysemy*). Figure 6 demonstrates that,

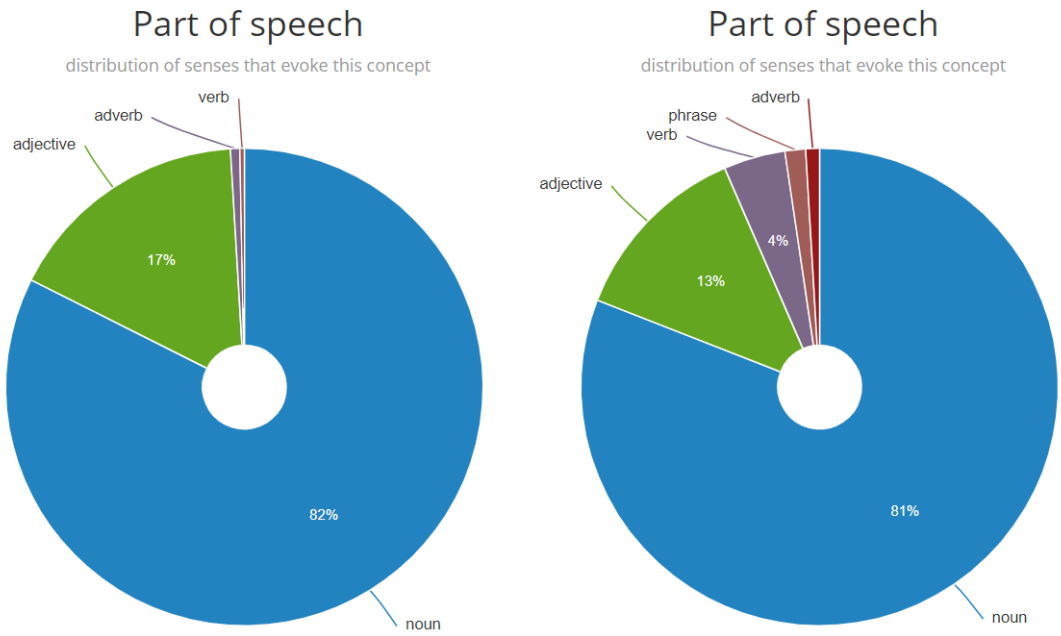


FIGURE 5 Distribution of Old English (left) and Old Frisian (right) senses in "02.03.02 Family/household"

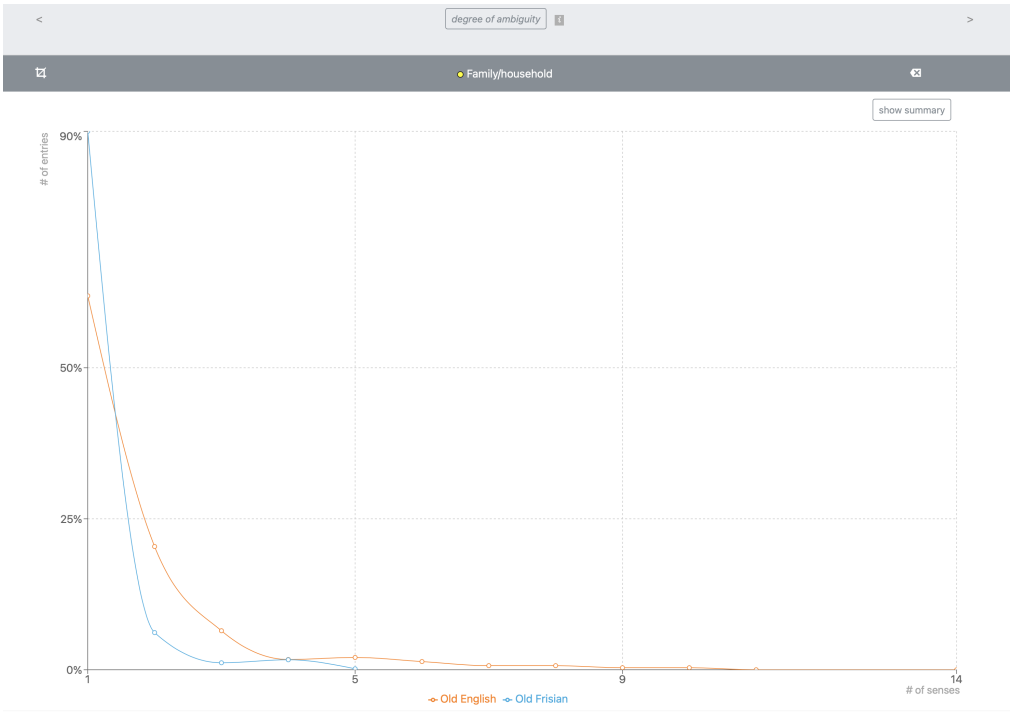


FIGURE 6 Degree of polysemy within "02.03.02 Family/household"

within the taxonomy branch of "02.03.02 Family/household", the vast majority of Old Frisian lemmata is monosemous (i.e., 90% has a single recorded sense), whereas Old English has, relatively speaking, more lemmata that are polysemous. This outcome can partially be explained by the fact that *AFWB*, which was used to obtain the Old Frisian lemmata and senses, is a concise dictionary

and therefore does not record senses extensively. Even so, *AFWB* records multiple senses for entries when these senses are distinct enough to be necessary for initial readings of Old Frisian texts. The lack of polysemy for Old Frisian is striking, even when the nature of the source dictionary is taken into account. Whether this finding is characteristic of the language itself remains as yet undecided. The apparent monosemous nature of Old Frisian may be due to the lack of register variety in the surviving corpus. The Old Frisian corpus is predominantly juridical in nature whereas the Old English one is much more balanced, containing samples of different style varieties and registers, resulting in a higher number of polysemous words.

277

9

6.3 *Onomasiological Distribution over Taxonomy Levels*

Figure 7 shows the distribution of lexical senses over the various levels of the taxonomy, which is another advanced analysis offered by Evoke.[18] This diagram indicates that Old Frisian has more recorded senses located at taxonomy levels with highly specialized meanings than Old English (see levels 8–12). Moreover, Old Frisian features senses that are allocated to levels beyond those in use for Old English (levels 10–12). Indeed, many of the categories newly created for the purposes of capturing KINSHIP in Old Frisian have been added as subordinate ones to *TOE* categories in the more specialized levels of the taxonomy. This diagram visualizes that outcome. A possible explanation may be that Old Frisian texts are mainly juridical in nature, very often pertaining to inheritance law, and therefore deal with more precise meanings that denote family relationships. A case in point is the degree of kinship, for which the Old Frisian lexis that has come down to us includes fine-grained senses (see also Table 2).

6.4 *Onomasiological Distribution over Categories*

Distributions over thesaurus categories yield data regarding the degrees of lexicalization (also known as cultural elaboration) of semantic fields, which enables comparisons between them (Wierzbicka, 1997: 10–11). Figure 8 charts such a distribution for the subcategories of "02.03.02 Family/household", generated with the advanced statistics section of Evoke.[19] The Y-axis has been

18 The diagram indicates a total of 221 Old Frisian senses for this semantic field rather than the 215 senses mentioned in section 5. The disparity lies in the fact that six senses from *AFWB* have been allocated to not one but two *TOE* categories, effectively creating two subsenses for each of these *AFWB* senses when assigning them to the onomasiological structure of *TOE* (see Appendix A).

19 The diagram indicates a total of 220 Old Frisian senses spread over the semantic subfields. One of the senses from the 221 senses that one would expect (see previous footnote) is not found in the subfields but is allocated to the category "02.03.02 Family/household" itself, which accounts for the disparity.

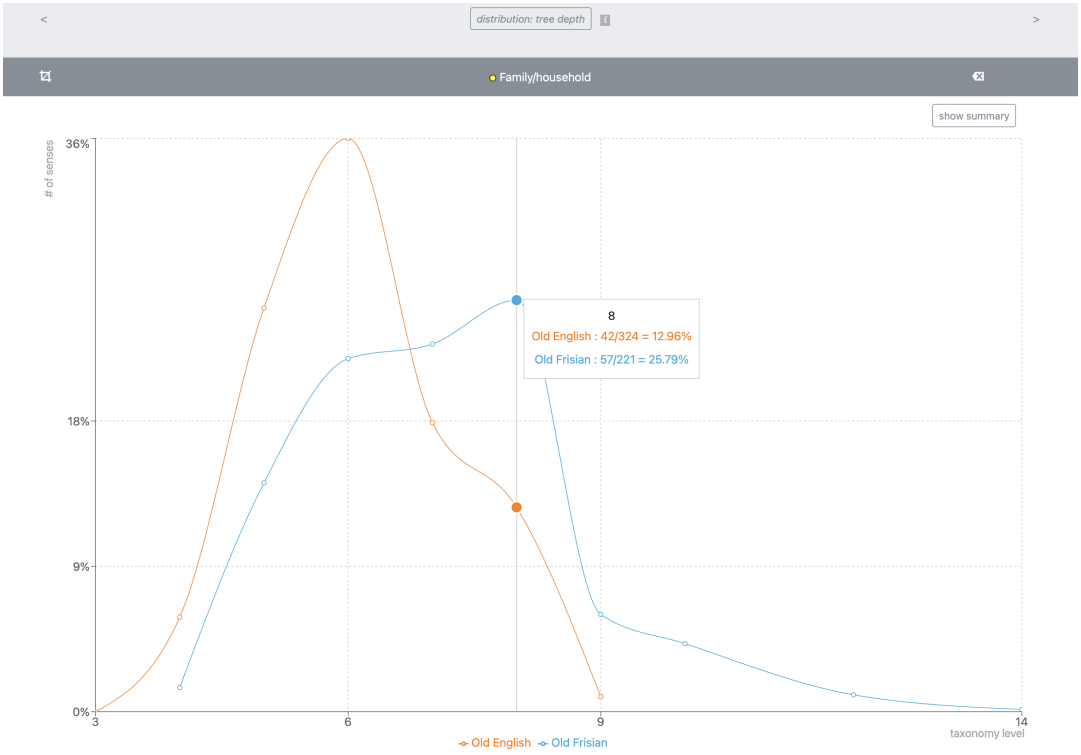


FIGURE 7 Distribution of lexical senses within "02.03.02 Family/household" over the taxonomy levels

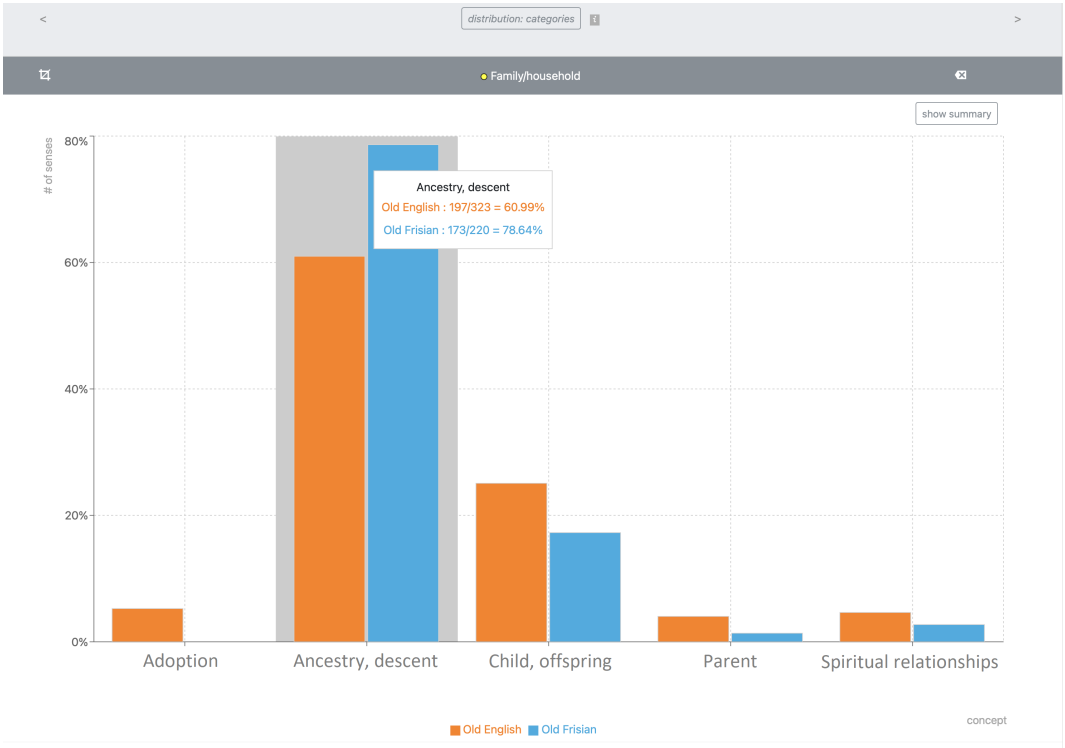


FIGURE 8 Distribution of lexical senses over the semantic subfields of "02.03.02 Family/household"

configured to show the relative number of senses from a single language (i.e. Old Frisian or Old English) found within each branch indicated on the X-axis. The branch "02.03.02.03 Ancestry, descent", highlighted in the diagram, contains the vast majority of the Old Frisian senses on KINSHIP (170 senses or 79%). The majority of Old English senses is found in the same branch, albeit less dominant (61%) in relation to the other branches within the field. In fact, "02.03.02.03 Ancestry, descent" is the sole branch for which Old Frisian has a higher relative number of senses recorded than Old English. All other branches – i.e. "02.03.02.04 Adoption", "02.03.02.02 Child, offspring", "02.03.02.01 Parent", and "02.03.02.05 Spiritual relationships" – have more Old English senses recorded than Old Frisian ones both in absolute and in relative numbers. The most striking differences between the two languages on this level are, therefore, (1) the relative degrees of lexicalization of "02.03.02.03 Ancestry, descent" and (2) the lack of any recorded Old Frisian senses for the concept of "02.03.02.04 Adoption".

Apart from "02.03.02.04 Adoption", the Old Frisian corpus does not contain words for a number of other concepts found in Old English. These concepts are, most notably, represented by the *TOE* categories of "02.03.02.04.01 Foster relationships", "02.03.02.02 | 06.01 A foundling", "02.03.02.02.01 Twins", and "02.03.02.02.02 Triplets".[20] KINSHIP concepts that witness a larger degree of lexicalization in Old Frisian in comparison to Old English are those that have been newly introduced (see Appendix B), of course. However, they also include concepts that are gender neutral (such as expressed with Old Frisian *swesterne* 'sibling', for which *TOE* records no Old English equivalent) and concepts represented in *TOE* by the categories "02.03.02.03.06.02.06 In-law relationships", "02.03.02.03.06.02.03 Child of brother/sister", "02.03.02.03.06.02.04 Cousin", and "02.03.02.02.05 | 02 A Bastard".[21]

20 "Foster relationships" in *TOE*. Identifier: 1268. Location: 02.03.02.04.01. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1268.

"A foundling" in *TOE*. Identifier: 1123. Location: 02.03.02.02/06.01. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1123.

"Twins" in *TOE*. Identifier: 1127. Location: 02.03.02.02.01. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1127.

"Triplets" in *TOE*. Identifier: 1130. Location: 02.03.02.02.02. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1130.

21 "In-law relationships" in *TOE*. Identifier: 1256. Location: 02.03.02.03.06.02.06. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1256.

"Child of brother/sister" in *TOE*. Identifier: 1239. Location: 02.03.02.03.06.02.03. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1239.

An extensive analysis of the distributions found in the more specific levels of the taxonomy branches is beyond the scope of this article. Nevertheless, to show what results such an analysis may produce, we include some insights into one such distribution here. Figure 9 presents the dispersion for "02.03.02.02.05 Having the same parents", a subcategory of "02.03.02.02 Child, offspring", which has a high degree of lexicalization for Old Frisian compared to Old English.[22] Some interesting observations can be made about this diagram: Old Frisian has more words than Old English with senses of "02.03.02.02.05.02 Sister" and "02.03.02.02.05 | 02 A Bastard". The latter is even expressed with a word specific to a child born before its parents were married: *spilkind*. The category "02.03.02.02.05.03 Siblings" has been created for the Old Frisian lexis, since no

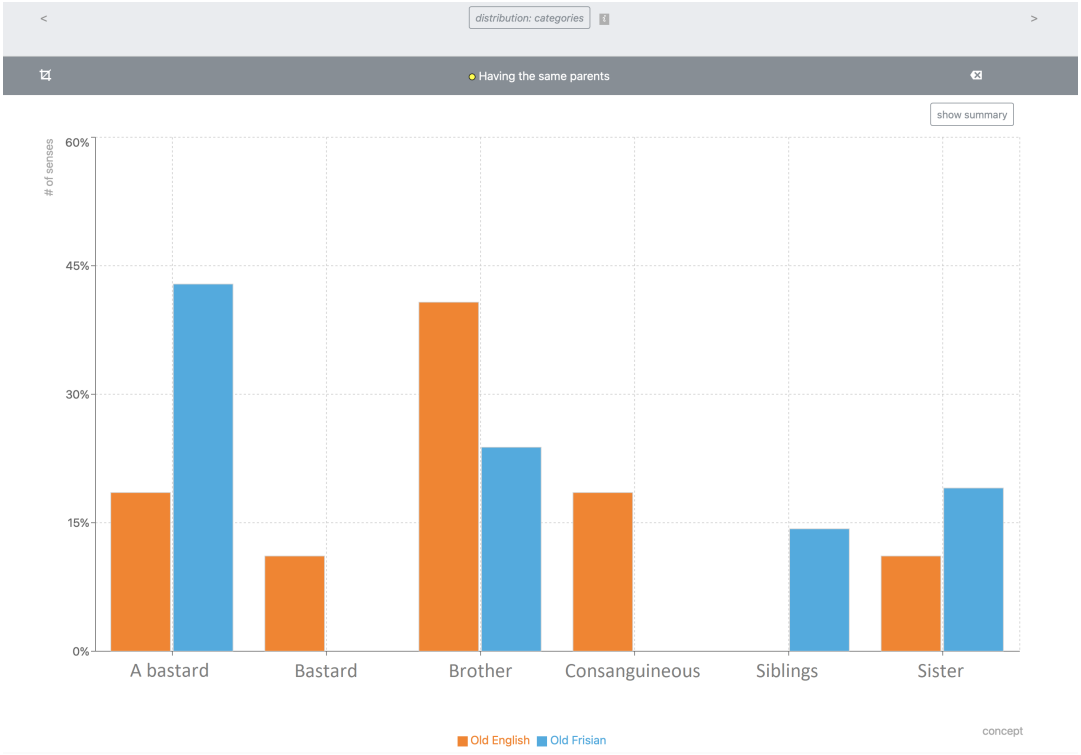


FIGURE 9 Distribution of lexical senses over the semantic subfields of "02.03.02.02.05 Having the same parents"

"Cousin" in *TOE*. Identifier: 1247. Location: 02.03.02.03.06.02.04. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1247.

"A bastard" in *TOE*. Identifier: 1141. Location: 02.03.02.02.05/02. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1141.

22 "Having the same parents" in *TOE*. Identifier: 1136. Location: 02.03.02.02.05. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1136.

"Child, offspring" in *TOE*. Identifier: 1113. Location: 02.03.02.02. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1113.

Old English lexemes are recorded for this concept that leaves gender unspecified. The higher degree of lexicalization of both "02.03.02.02.05.02 Sister" and "02.03.02.02.05.03 Siblings" in Old Frisian compared to Old English, along with a lower degree for "02.03.02.02.05.01 Brother", suggests that the level of expressivity for this kinship tie is more alike for members of the male and female sex in medieval Frisia than in the Anglo-Saxon kingdoms. Further research is warranted into the question whether this hypothesis will hold when these semantic fields are compared for attestation of lexis in solely juridical texts, which constitute the majority of the surviving Old Frisian written legacy but only a fraction of the much vaster Old English corpus.

281

9

7 Discussion

The analyses and results in the previous sections are to be read in the context of the languages and resources that lie at their heart. Old Frisian and Old English are not contemporaneous languages: the surviving sources for Old Frisian are coeval with the period of Middle English. The observed contrasts in comparing these languages, however similar they may be, will therefore likely be influenced by the temporal as well as regional space between them. Likewise, it is important to bear in mind the nature of the corpora from which the lexicon was reconstructed. Surviving texts represent but a small portion of what must have been written, by a non-homogeneous group, and, perhaps more importantly, solely by those who were literate. Religious and administrative texts therefore represent a large portion of these medieval corpora, with certain genres more dominant than others (e.g., homilies in Old English, legal documents in Old Frisian).

The alignment of Old Frisian senses with the semantic hierarchy of *TOE* was complicated by differences between the lexicographic resources used (i.e. *AFWB* for Old Frisian and *TOE* for Old English) and the cultural contexts in which they were created. *AFWB* and *TOE* use different languages and practices to describe their lexicon: the former employs Modern German to define senses, the latter Modern English; *AFWB* is a concise dictionary; *TOE* is based on more detailed dictionaries and demands sense differentiation to be of use. Allocating senses from one language to a taxonomy of a resource created for another, then, is by no means straightforward (see Appendix A for notes). As a result, observations with lingual comparisons, such as those made in this article, reflect differences between not only the language communities concerned, but also between the lexicographic practices that contributed to the frameworks used for interpretation of the lexis.

8 Conclusion

In this study we set out to answer two questions. The first is whether it is possible to allocate the Old Frisian lexis within the semantic field of KINSHIP to the onomasiological macrostructure of *TOE*. The answer is in the affirmative. We have demonstrated that Old Frisian senses for KINSHIP can be viewed in an onomasiological structure, alongside Old English ones, by reusing the *TOE* macrostructure. However, the process of allocating senses from one language to a taxonomy of a resource created for another is by no means straightforward, as mentioned before. In addition to differences between the lexicographic practices for the two resources that have been aligned, a substantial number of Old Frisian senses, owing to their specialized meaning, demanded new categories to be fashioned and positioned into the taxonomy of *TOE*. For the domain of KINSHIP, these newly created categories could be slotted into lower, more specialized levels of the semantic hierarchy of *TOE*. The current research does not yet allow us to establish whether reuse and extension of an existing onomasiological structure was more time efficient than building one from the ground up. Of course, creating a new hierarchy, rather than reusing that of *TOE*, would have the disadvantage of forestalling onomasiological comparisons between Old Frisian and Old English. We surmise that adoption of semi-automated approaches (e.g., automated recognition of cognates) may be used in the future to significantly speed up the alignment process.

The second question that we have aimed to answer is whether Evoke, in combination with *TOE*, can offer new insights both for Old Frisian and, in contrast to Old Frisian, Old English. As demonstrated, there are a number of advantages to having Old Frisian lexis available in the onomasiological structure of a thesaurus. The first is that the resulting resource facilitates word field studies (comparable to those for which *TOE* has been used in the context of Old English) and comparative linguistic research (see the Results section). In fact, we expect the Old Frisian lexis to be accessible to a larger audience through Evoke, owing to the availability of Old Frisian senses in a digital resource that contains Modern English headings, using the *TOE* macrostructure, rather than in a dictionary that records sense definitions in German. A second advantage is that statistical analyses such as those enabled by Evoke lead to new knowledge of Old Frisian lexis. Preliminary analyses have already demonstrated that the field of KINSHIP in the surviving Old Frisian lexis consists of significantly fewer adjectives and more verbs compared to Old English; it contains lemmata that are mostly monosemous (90%); it includes more fine-grained senses than Old English (including ones to denote different degrees of kinship); it has a relatively higher degree of lexicalization of the concepts of ancestry and

descent than Old English; but it lacks any words for the concepts of adoption, foundling, twins, and triplets. Findings in Evoke lead to new questions that merit further research – into the surviving corpus and lexicographic practices, amongst others – to supply a satisfying context and better understanding. The availability of both Old Frisian and Old English lexis in Evoke, then, certainly offers a useful stepping stone to learn more about the nature of these kindred historical languages and their language communities.

References

AFWB = Hofmann, D., and A. T. Popkema. *Altfriesisches Handwörterbuch* (Heidelberg: Universitätsverlag Winter, 2008).

Bajema, I. M. "The Mother's Brother: An Investigation into the Meaning of Old English *eam*." *Neophilologus* 78 (1994), 633–643.

Baker, P. S. *Introduction to Old English* (Chichester: Wiley-Blackwell, 2012).

Baldi, P. *Linguistic Change and Reconstruction Methodology* (Berlin and New York: De Gruyter Mouton, 2011).

Bammesberger, A. "Altfriesisch *swâger*." *Indogermanische Forschungen* 73 (1968), 133–135.

Boersma, U. J. "De frou yn 'e Fryske wetten." *Us Wurk* 10 (1961), 58–85.

Boutkan, D., and S. Siebinga. *Old Frisian Etymological Dictionary* (Leiden: Brill, 2005).

Bremmer Jr, R. H. "Insults Hurt: Verbal Injury in Late Medieval Frisia." In *Approaches to Old Frisian Philology*, eds. R. H. Bremmer Jr, T. S. B. Johnston, and O. Vries (Amsterdam: Rodopi, 1998), 89–112.

Bremmer Jr, R. H. "Language and Contents of the Old Frisian Manuscripts from Rüstringen (c.1300): A 'Veritable *Mixtum Compositum*'." In *Advances in Old Frisian Philology*, eds. R. H. Bremmer Jr, S. Laker, and O. Vries (Amsterdam: Rodopi, 2007), 29–64.

Bremmer Jr, R. H. "Taking Stock of Old Frisian Studies: 1992–2021." *Us Wurk* (2021), 1–28.

Bremmer Jr, R. H. "The Importance of Kinship: Uncle and Nephew in *Beowulf*." *Amsterdamer Beiträge zur älteren Germanistik* 15 (1980), 21–38.

Bremmer Jr, R. H. *A Bibliographical Guide to Old Frisian Studies* (Odense: Odense UP, 1992).

Bremmer Jr, R. H. *An Introduction to Old Frisian: History, Grammar, Reader, Glossary* (Amsterdam: John Benjamins Publishing Company, 2009).

Chandler, D., and R. Munday. *A Dictionary of Media & Communication*, 2nd edn (Oxford: Oxford University Press, 2016).

Colleran, R. A. B. *Keeping It in the Family. Disentangling Contact and Inheritance in Closely Related Languages*. Doctoral Dissertation (The University of Edinburgh, 2017).

Dance, R. Review of *TOE*, *Medium Ævum* 66 (2) (1997), 312–313.

Díaz-Vera, J. E. "Reconstructing the Old English Cultural Model for Fear." *Atlantis* 33 (2011), 85–103.

Diepeveen, A. "Verwantschapstermen in de Germaanse talen." *Antwerp Papers in Linguistics* 105 (Universiteit Antwerpen, 2003).

Durkin, P. "New Light on Early Middle English Borrowing from Anglo-Norman: Investigating Kinship Terms in *grand-*." *Anglia* 137 (2019), 255–277.

DWDS = Digitales Wörterbuch der deutschen Sprache (Berlin: Berlin-Brandenburgischen Akademie der Wissenschaften), https://www.dwds.de/d/wb-dwdswb.

Evoke = Stolk, S. Evoke (Web application, 2018), http://evoke.ullet.net/.

Fischer, A. "Of *fæderan* and *eamas*: Avuncularity in Old English." In *The Power of Words: Essays in Lexicography, Lexicology and Semantics in Honour of Christian J. Kay*, eds. G. D. Caie, C. Hough and I. Wotherspoon (Amsterdam: Rodopi, 2006), 67–77.

Fox, R. *Kinship and Marriage: An Anthropological Perspective*, 2nd edn (Cambridge: Cambridge UP, 1984).

Friel, B. M., and S. M. Kennison. "Identifying German-English Cognates, False Cognates, and Non-Cognates: Methodological Issues and Descriptive Norms." *Bilingualism: Language and Cognition* 4 (3) (2001), 249–74.

Görlach, M. Review of *TOE*, *Anglia* 116 (3) (1998), 398–401.

Hofmann, G., ed. *Mittelalterliches Alltagsleben im Spiegel der altfriesischen Terminologie: mit Ergänzungen aus zeitgenössischen niederländischen Quellen*, Estrikken 98 (Kiel: Nordfriesische Wörterbuchstelle Christian-Albrechts-Universität, 2015).

Hough, C. "Old English *weargbeorg*." *Notes and Queries* 54 (4) (2007), 364–365.

Kok, W. "Houlik en earberens by us forfaers." *It Beaken* 9 (1947), 59–64, 67–73.

Lancaster, L. "Kinship in Anglo-Saxon Society: I." *The British Journal of Sociology* 9 (1958), 230–250.

Lancaster, L. "Kinship in Anglo-Saxon Society: II." *The British Journal of Sociology* 9 (1958), 359–377.

Lowe, K. A. "Never Say *nefa* Again: Problems of Translation in Old English Charters." *Neuphilologische Mitteilungen* 94 (1993), 27–35.

Loyn, H. R. "Kinship in Anglo-Saxon England." *Anglo-Saxon England* 3 (1974), 197–209.

Meijering, H. D. *Chindh uuirdit uns chiboran. Over het woord 'kind' in het oudere Duits* (Amsterdam: Vrije Universiteit, 1985).

Meijers, E. M. "Eine germanische Zählung der Verwandtschaftsgrade." *Tijdschrift voor Rechtsgeschiedenis* 6 (1925), 1–52.

Munske, H. H. *Der germanische Rechtswortschatz im Bereich der Missetaten: Philologische und sprachgeographische Untersuchungen. I: Die Terminologie der älteren westgermanischen Rechtsquellen* (Berlin: De Gruyter, 1973).

Munske, H. H. et al., eds. *Handbuch des Friesischen. Handbook of Frisian Studies* (Tübingen: Max Niemeyer Verlag, 2001).

Ojanguren López, A. E. "Old English Try Verbs: Grammatical Behaviour and Class Membership." *Studia Neophilologica* (2021), 1–19.

Otwinowska-Kasztelanic, A. "Awareness and Affordances: Multilinguals Versus Bilinguals and Their Perceptions of Cognates." In *New Trends in Crosslinguistic Influence and Multilingualism Research*, eds. G. De Angelis and J. M. Dewaele (Bristol and Blue Ridge Summit: Multilingual Matters, 2011), 1–18.

Pfeffer, G. "The Vocabulary of Anglo-Saxon Kinship." *L'Homme* 27 (1987), 113–128.

Popkema, A. T. "Die altfriesischen Eidesbezeichnungen." In *Advances in Old Frisian Philology*, eds. R. H. Bremmer Jr, S. Laker, and O. Vries (Amsterdam: Rodopi, 2007), 325–356.

Pospelova, K. "Old Frisian Compounds: Kinship Terms." *Žmogus ir Žodis* 19 (2018), 79–96.

Rauer, C. "*Mann* and Gender in Old English Prose: A Pilot Study." *Neophilologus* 101 (1) (2017), 139–158.

Roberts, J. "Towards an Old English Thesaurus." *Poetica* 9 (1978), 56–72.

Salmons, J. "What Old Frisian Can Tell Us about the History of i-Umlaut across West Germanic?" In *Advances in Old Frisian Philology*, eds. R. H. Bremmer Jr, S. Laker, and O. Vries (Amsterdam: Rodopi, 2007), 367–377.

Schmitt, N. "Vocabulary Learning Strategies." In *Vocabulary: Description, Acquisition and Pedagogy*, eds. N. Schmitt and M. McCarthy (Cambridge: Cambridge UP, 1997), 199–227.

Scott, P. "Conceptualising Olfaction: A Study of the Scent Nouns and Adjectives in Old English." *Studia Neophilologica* (2021), 1–18.

Sjölin, B. "Zur Gliederung des Altfriesischen." *Us Wurk* 15 (1966), 25–38.

Spolsky, E. "Old English Kinship Terms and *Beowulf*." *Neophilologische Mitteilungen* 78 (1977), 233–238.

Sterringa, A. "The Most Miserable Women of All: Widows in Medieval Frisia." In *Approaches to Old Frisian Philology*, eds. R. H. Bremmer Jr, T. S. B. Johnston, and O. Vries (Amsterdam: Rodopi, 1998), 285–301.

Stroh, F. *Handbuch der germanischen Philologie* (Berlin: De Gruyter, 1985).

Tissari, H. "Current Emotion Research in English Linguistics: Words for Emotions in the History of English." *Emotion Review* 9 (1) (2017), 86–94.

TOE = Roberts, J., and C. Kay with L. Grundy. *A Thesaurus of Old English* (Glasgow: University of Glasgow, 2017), http://oldenglishthesaurus.arts.gla.ac.uk/.

Trask, R. L. *Trask's Historical Linguistics*, ed. R. McColl Millar. 3rd edn (London and New York: Routledge, 2015).

Van de Poel, R. *Corpus Oudfries* (2019), http://corpora.ato.ivdnt.org/corpus-frontend/OFR/.

Versloot, A. P. "Mechanisms of Language Change: Vowel Reduction in 15th Century West Frisian." Doctoral dissertation (University of Groningen, 2008).

285

9

Versloot, A. P. "Why Old Frisian Is Still Quite Old." *Folia Linguistica Historica* 38 (2004), 253–298.

Vries, O. "Die altfriesischen Urkunden." In *Handbuch des Friesischen. Handbook of Frisian Studies*, eds. H. H. Munske et al. (Tübingen: Max Niemeyer Verlag, 2001), 594–601.

Wierzbicka, A. *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese* (Oxford: Oxford University Press, 1997).

Zeilfelder, S. "Das Wortfeld 'Mensch'." In *Deutsche Wortfeldetymologie in europäischem Kontext, Band 3: Mensch und Mitmensch*, ed. R. Lühr (Wiesbaden: Reichert, 2015), 33–110.

Appendix A: Notes on the Alignment

This appendix provides notes on the alignment between the *Altfriesisches Handwörterbuch* (AFWB) and *A Thesaurus of Old English* (*TOE*) for the semantic field of KINSHIP.

A.1 AFWB *Lemmata or Senses Discarded or Updated*

- The second sense recorded in *AFWB* for *evenknē*, i.e. "Verwandter" [relative], is disputable and has therefore been discarded.
- Senses recorded in *AFWB* for *bern* and *kind* do not reflect that they are synonyms. With the help of the online corpus of Old Frisian, it is easy to find instances that attest to the conclusion that *bern* and *kind* have the same senses, i.e. 1) Foetus, 2) Offspring (first degree), 3) Descendant and 4) Young, immature person. We have therefore chosen to align the senses of *bern* and *kind* with each other.
- The recorded sense in *AFWB* for *unnatlik*, i.e. "unebenbürtig" [not equally related], is inaccurate and should be "not related."
- The senses of Old Frisian *kennemech* has been aligned with those of *keremech*.

A.2 *Lemmata or Senses Introduced*

The following terms were not recorded in *AFWB* but are attested in the Old Frisian corpus and were added to the dataset used in this article.

- *sex honda* as a phrase with the sense "First degree blood relatives".
- *thredda* as a noun with the sense "Third generation".

The phrase s*ex honda* is found in 7 manuscripts, in one of the most widely distributed texts known as the *Twenty-four Land Laws*.[23] *Thredda* is recorded in the *AFWB* as

23 This sense is attested in the following Old Frisian text witnesses: PnB-E3, L24-H2, L24-J, L24-R1, L24-E1, L24-F, JF-Ro. See results of a search for "sex hond" in *Corpus Oudfries*.

numerical form with the general sense of *dritte* [third], but the more specialized sense of "third generation" is also recorded in the Old Frisian sources.[24]

A.3 *Remarks on Placement of Senses in* TOE

– Definitions in *AFWB* of a number of Old Frisian senses on KINSHIP contained elements that made it possible to allocate these senses to fields outside of KINSHIP, too. Instead of allocating them to multiple semantic fields, these senses were allocated solely to categories in KINSHIP. The Old Frisian lemma *kniaia*, for instance, has a sense of "lawfully claiming to be related". This sense was allocated to KINSHIP in *TOE* only, but not additionally to the semantic field of "14.03.03 n. Law, action of the courts".[25] Similarly, lemmata with the sense of "being related" or "relationship" were not allocated to the semantic field of "08.01 n. Heart, spirit, mood, disposition" in *TOE*.[26]
– The *AFWB* lemma *kniaia* has three overlapping senses: "Verwandschaft nachweisen," "Verwandschaft geltend machen," "verwandt sein." We merged these three into one sense in English: "To (claim to) be related."
– Senses defined as "Blutverwandt" or "Verwandt" are used indiscriminately in the *AFWB*. These senses have all been aligned as "Related." Similarly, the *AFWB* uses "Blutverwandtschaft" and "Verwandtschaft" indiscriminately. Blood relations and blood relatives all belong to *TOE* category "Close relationship."

A.4 *Comments on the Structure of* TOE

In the following paragraphs we illustrate some of the difficulties experienced when placing the Old Frisian senses from *AFWB* within the categories from *TOE*.

– The *AFWB* records a single sense for the Old Frisian lemma *dochter* with its definition in German: "Tochter" [daughter]. *TOE* records the Old English cognate *dohtor* in different senses: one is listed in the category "02.03.02.02.04 A daughter", another in the category "02.03.02.03.05 | 02 n. Female descendant".[27] The Old Frisian lemma's meaning refers to both of these. The question arises as to whether one should allocate the single recorded sense of Old Frisian *dochter* to the same two categories as

287

9

24 This sense is attested in the following Old Frisian text witnesses: LaFi-F, PnB-B2, Dom-J, BHuB, Lav-H2, LaFi-Ro. See results of a search for "thredda" in *Corpus Oudfries*.

25 "Law, action of the courts" in *TOE*. Identifier: 19453. Location: 14.03.03. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=19453.

26 "Heart, spirit, mood, disposition" in *TOE*. Identifier: 13458. Location: 08.01. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=13458.

27 "A daughter" in *TOE*. Identifier: 1133. Location: 02.03.02.02.04. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1133.

"Female descendant" in *TOE*. Identifier: 1193. Location: 02.03.02.03.05/02. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1193.

Old English *dohtor*. Since it is obvious that a daughter is also a female descendant, a sense that is also part of KINSHIP, we have decided to allocate the recorded sense in *AFWB* to both categories. An alternative solution would be to create two subsenses of the *AFWB* lemma and have allocated these to their respective categories.

– The senses of the cognate words Old English *fǣmne* and Old Frisian *fomne* are related but still slightly different (see Table A1).

TABLE A1 *fǣmne* and *fomne*

Lemma	***TOE* categories**		
Old English *fǣmne*	Female person, woman	Girl	A maiden, A virgin
	***AFWB* senses**		
Old Frisian *fomne*	1) Mädchen [girl]	2) Tochter [daughter]	3) Magd [female servant]

When looking at the *TOE* categories for *fǣmne*, the category "02.03.01.02 Female person, woman" is a sense more general than those recorded in *AFWB*.[28] Although the lemma may fit in this category, too, we have decided to keep the three senses recorded in *AFWB* unchallenged, positioning only these in the semantic hierarchy of *TOE* on KINSHIP.

A.5 *Inconsistencies* AFWB

The senses of (near-)synonyms and other semantically related words have some inconsistencies. Cases in point are Old Frisian *wīf* and *frouwe*.

> **wīf** n. 1) Weib; 2) Frau
> **frouwe, frowe, frou**^{WL} f. 1) Frau; 2) Ehefrau; 3) Edelfrau, Fürstin; 4) Herrin; 5) die Jungfrau Maria; 6) Schwiegermutter

Examination of the contexts in which the lexemes *wīf* and *frouwe* occur reveals that the compilers of the *AFWB* have tried to reflect the connotational differences that the two words have in the German descriptions. *Frouwe* was a term used to denote respect, while *wīf* was the more generally used word. For *frouwe* only the senses 1, 2

28 "Female person, woman" in *TOE*. Identifier: 1079. Location: 02.03.01.02. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1079.

and 6 have been allocated to KINSHIP in *TOE* (the others fall outside the scope of our investigation). Corresponding *TOE* categories are "02.03.01.02 Female person, woman" and "12.09 | 07.05 n. Wife, married woman."[29]

A.6 *Allocation of* AFWB *Senses to Multiple* TOE *Categories*

Six senses from *AFWB* have been allocated to not one but two *TOE* categories, effectively creating two subsenses for each of these *AFWB* senses when assigning them to the onomasiological structure of *TOE*. The allocation to a second *TOE* category allowed, most notably, for words denoting offspring (such as "son" or "daughter") to be marked as "descendant", too. The six *AFWB* senses, along with their allocations in *TOE*, are in Table A2.

TABLE A2 Six senses

AFWB				*TOE*	
Lemma	**Sense no.**	**Sense**	**Mod. English translation**	**Category 1**	**Category 2**
dochter	1	Tochter	daughter	A daughter[a]	Descendant[b]
kind	2	Kind, Nachkomme	child, descendant	Child, offspring[c]	Descendant
knapa	2	Sohn	son	A son[d]	Descendant
sunder_1	1	Sohn	son	A son	Descendant
sune	1	Sohn	son	A son	Descendant
slachte_1	2	Familie	family	Ancestry[e]	Kinsman/relative[f]

a "A daughter" in *TOE*. Identifier: 1133. Location: 02.03.02.02.04. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1133.

b "Descendant" in *TOE*. Identifier: 1191. Location: 02.03.02.03.05. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1191.

c "Child, offspring" in *TOE*. Identifier: 1113. Location: 02.03.02.02. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1113.

d "A son" in *TOE*. Identifier: 1131. Location: 02.03.02.02.03. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1131.

e "Ancestry" in *TOE*. Identifier: 1159. Location: 02.03.02.03. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1159.

f "Kinsman/relative" in *TOE*. Identifier: 1208. Location: 02.03.02.03.06.01. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1208.

29 "Wife, married woman" in *TOE*. Identifier: 18618. Location: 12.09/07.05. IRI: http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=18618.

Appendix B: Old Frisian Lemmata on Kinship Aligned with *A Thesaurus of Old English*

Categories in italics are ones that did not exist in the original thesaurus taxonomy and have been newly created to accommodate Old Frisian senses.

02.03.02 **Family/household (id: 1108)**
IRI: https://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1108

02.03.02.01 **Parent (1109)**	alder
01 n. Father (1110)	feder
02 n. Mother (1112)	mōder
02.03.02.02 **Child, offspring (1113)**	bern, kind
04 adj. Lawfully born (1120)	aftberen
06 n. An orphan (1124)	wēsa, wēsekin, wēsekind, wēseklēn, wēstien
02 adj. Without Parents (1126)	alderlās, biwēsed, wēsid
05 adj. For orphans	wēslik
02.03.02.02.03 **n. A son (1131)**	knapa, sunder, sune
02.03.02.02.04 **n. A daughter (1133)**	fomne, dochter
02.03.02.02.05 **Having the same parents (1136)**	
02 n. A Bastard (1141)	basterd, hōringe, hōrkind, hōrning, hōrbrēd, hōrbern
02 n. Bastard brother (1143)	basterdbrōther
03 n. Bastard sister	basterdswester
04 n. Bastard born before marriage	spilkind
02.03.02.02.05.01 **Brother (1146)**	brōther
06 adj. Of brothers, brotherly (1154)	brōtherlik
06.01 n. State of being brotherly	brōtherlikhēd
08 n. Full brother	fulbrōther
09 n. Half brother	halfbrōther
02.03.02.02.05.02 **Sister (1156)**	swester
01 n. Full sister	fulswester
02 n. Half sister	halfswester
02 n. Sisterhood	swesterskip
02.03.02.02.05.03 ***Siblings***	swesterne
01 n. Full siblings	fulswesterne
02 n. Half siblings	halfswesterne

02.03.02.03 Ancestry, descent (1159)	slachte
02 Side (1162)	sīde
03 Generation, degree of descent (1163)	grād, gung, knī, lith, sibdēl, siā, slachte, sibdēl
02 Third generation	thredda
03 Descendance	delgung, nithergung, tōdele
01 v. To descend	delgunga, nithergunga, tōdelgunga
02 adj. Descending, down	delgungande
03 adv. Descendingly	tōdele
04 Ascendance	upgung
01 v. To Ascend	upgunga, ūrbekgunga
02 adv. Ascendingly, backwards, up	ūrbek
05 Part of a family tree	facht
06 Trunk of a family tree	stipa
02.03.02.03.01 Ancestry, paternal kinship (1166)	
02 Male line of descent (1169)	sperehond, swerdsīde
03 Father's side (1170)	federsīde
01 adj. Paternal (1171)	fetherlik
02 n. Father's mother side	federesmōdersīde
02.03.02.03.02 Maternal descent (1172)	
02 Female line of descent (1174)	spindelsīde
03 Mother's side (1175)	mōdersīde
01 adj. Maternal (1176)	mōderlik
04 n. Mother's father side	mōderfedersīde
02.03.02.03.03 Forefather, ancestor (1178)	alder, forefeder, forefirdera
02.03.02.03.04 (Of degrees of descent) great-, grand- (1182)	
01 n. Grandfather (1183)	aldafeder, alder, edela
02 n. Grandmother (1184)	aldemōder,
03 n. Great grandfather (1185)	ūraldafeder, alder, edela
04 n. Great grandmother (1186)	ūraldemōder
09 n. Granduncle	aldaēm
02.03.02.03.05 Descendant (1191)	bern, kind, knapa, neikuma, neikumande, neikumeling, neikumer, sunder, sune

291

9

01 n. Descendants (of a generation) (1192)	neikuminge
02 n. Female descendant (1193)	dochter
03 n. Successor, heir (1194)	*outside of current scope*
04 n. Grandson (1195)	-
01 n. Daughter's son (1196)	dochtersune
12 n. Grandchild	bernesbern, kindeskind
01 n. Daughter's child	dochteresbern
13 n. Great grandchild	kindeskindeskind
02.03.02.03.06 Kinship, relationship (1204)	sibbe_1, blōd, wirtel
01 adj. Related (1205)	bisib, bisibbed, sibbe_2, swēs, liāf, nātlik
02.03.02.03.06.01 Kinsman, relative (1208)	sibba, sibdēl, kniā, knīling, mēch, friund, nāt, holda
03 n. Paternal relative (1212)	federfriund, federmēch, federsibbe
04 n. Maternal relative (1213)	mōderfriund, mōdermēch
07 n. Next of kin (1216)	nesta
09 n. Kinsfolk (1218)	sibbe_1, slachte, ken
05 n.Relationship through common parents	fulsibbe
01 adj. Rel. through common parents	fulsibbe
06 n. Relationship by one parent	halfsibbe
11 n. Third degree relative	thredkniā, thredknīling, thredling, thredsiā
12 n. The six first degree relatives	sex honden
13 n. Relative in the same degree	evenknē, evenkniā, evenknīling, līkenisse
14 n. Relative chosen for oath swearing	keremēch, kestfriūnd
15 n. Relative in ascending line	upstīger
16 n. Relative in descending line	nītherstīger
02.03.02.03.06.02 Close relationship (1224)	
01 adj. Closely related (1225)	biefta, inlik, niār
02 adj. Like relatives	swēslik
03 n. Blood relations (1230)	blōd, bērena blōd
04 adj. Not related (1233)	fremede, unnātlik
05 v. To (claim to) be related	kniāia
01 v. to (claim to) be closer related	bikniāia, ūrkniāia, ūtsteka
06 adj. Nearest related	allerswesest, allersibbest, nest

07 adj. Too close related (*incest*)	sibbe_2
08 adj. More distantly related	ūtlik
02.03.02.03.06.02.01 Uncle (esp. maternal) (1234)	ēm, mōderesbrōther
01 n. Paternal uncle (1235)	federia
01 n. Father's halfbrother	halffederia
02.03.02.03.06.02.02 Aunt (1236)	
01 n. Maternal Aunt (1237)	mōdire, mōie
02 n. Paternal Aunt (1238)	fethe, federesswester
02.03.02.03.06.02.03 Child of brother/sister (1239)	swesternabern
01 n. Sister's child (1240)	swesterbern, swesterling, swesterkind
02 n. Nephew (1241)	neva
01 n. Sister's son (1242)	swestersune
02 n. Brother's son (1243)	brōthersune, neva
03 n. Niece (1244)	nifte, nifke
01 n. Brother's daughter (1245)	brōtheresdochter
02 n. Sister's daughter	swesterdochter
03 n. Niece's child	niftlin, niftakind
05 n. Brother's child	brōtheresbern, brōthereskind
02.03.02.03.06.02.04 Cousin (1247)	
02 n. Father's sister's son (1249)	fethansune
04 n Relationship between cousins	berning, swire
05 n. Mother's brother's son	ēmka, ēmessune
06 n. Mother's sister son	mōdiransune
07 n. Father's brother's son	federiasune
08 n. Children of cousin	efterswesterling, efterswesternabern, ōtherswesterbern
02.03.02.03.06.02.05 Step relationships (1251)	
01 n. Step-father (1252)	stiāpfeder
02 n. Step-mother (1253)	stiāpmōder
03 n. Step-son (1254)	stiāpsune
04 n. Step-daughter (1255)	stiāpdochter
05 n. Step-child	stiāpkind

293

9

02.03.02.03.06.02.06 In-law relationships (1256)	
01 n. Father-in-law (1257)	hēra, swāger
02 n. Mother-in-law (1258)	frouwe
03 n. Daughter-in-law (1259)	snore
01 n. Being a daughter-in-law	snōrskip
04 n. Brother-in-law (1260)	tāker, āthum
06 n. Son-in-law/brother-in-law (1262)	āthum
07 n. Son-in-law and father-in-law (1263)	siāring
08 n. Son-in-law	swiāring, swāger, dochtermon
01 n. Being a son-in-law	swāgerskip
09 n. Parents-in-law	swiāring
02.03.02.04 Adoption (1264)	-
02.03.02.04.01 Foster relationships (1268)	-
02.03.02.05 Spiritual relationships (1274)	fadersibbe, faderskip
01 adj. Spiritual (1275)	gāstlik
02.01 n. A godfather (1277)	fadera
05 n. Relationship betw. godbrother and biological brother	brōtherlikhēd
06 n. Relationship betw. godfather and biological father	fēderlikhēd

B.1 *Other* TOE *Categories Containing Old Frisian Lemmata*

As mentioned in section 5, a number of Old Frisian lemmata on KINSHIP have been classified within *TOE* outside of the semantic field "02.03.02 Family/household", and are therefore not included in our current analyses and case studies. The classification of these 47 lemmata is provided below.

02.01 Existence, life (id: 661)
IRI: https://oldenglishthesaurus.arts.gla.ac.uk/category/#id=661

02.01.03 Fruitfulness, fertility (698)	
02.01.03.05 To be fruitful, to produce (705)	forthbrenga
02.01.03.02 Barrenness, sterility (717)	
02 adj. Barren, unproductive (719)	tochtalās

02.01.03.03.01 To beget (730)	
02 n. Begetting of children (732)	berntām, berntochta
06 n. An embryo/foetus (738)	berde, bern, kind
09 adj. Pregnant (742)	bernheftich
12 adj. Childless (745)	bernlās
02.01.03.03.02 To bring forth, produce (747)	
05 n. Produce, fruit, offspring (752)	frucht
02.01.03.03.03 Birth (761)	berde
01 v. To be born (762)	bera_1
02 adj. Born (763)	beren
02.01.03.03.03.01 Child-bearing, childbirth (768)	tām
07 n. A miscarriage (776)	bernwendene, berdwendene, dāden frucht
02.01.03.03.04 Offspring, race, breed, family, children (780)	*not included*
01 n. Progeny, offspring (781)	tām
02 n. Descendants, progeny, race (784)	*not included*
01.01.03.03.06 Sex, kind (789)	
01 n. A sex (790)	
01.01.01 n. A Male (793)	mon
02.01 n. A Female (795)	frouwe, wīf
02.03 Humankind (id: 1059)	geslachte, slachte_1

IRI: https://oldenglishthesaurus.arts.gla.ac.uk/category/#id=1059

02.03.01 People (1065)	menneska
01 n. Human being (1066)	mon, menneska
09 v. To have (a brother, husband etc.) (1075)	habba
02.03.01.01 Male person, man (1076)	mon, hēra, monnesnoma, monnespersōne
02 adv. In a male manner (1078)	monlike

02.03.01.02 Female person, woman (1079)	frouwe, wīf, wīfke, wīfnama, menneska, frouwespersōna
03 adj. Female (1082)	frouwelik, wīflik
02.03.01.04 Child (1086)	kind, kintien, mach, bern
05 n. Male child (1091)	knapabern, knapakind
06 n. Female child (1092)	fomnabern, fomnakind
02.03.01.05 Youth, boy, stripling (1093)	knapa, mage
02.03.01.07 Girl (1099)	famke, fomne, megeth, meiden

12.09 Marriage, state of marriage (id: 18602)
IRI: https://oldenglishthesaurus.arts.gla.ac.uk/category/#id=18602

06 n. Married persons, married couple (18610)	hiūne, sinhīgen
07.03 n. Husband, (married) man (18615)	mon
07.05 n. Wife, married woman (18618)	wīf, wīfke, wīfnama, hūsfrouwe
12.09.03 Unmarried state (18657)	
02 A maiden, a virgin (18663)	megeth
03 An unmarried man (18665)	knapa
12.09.05 State of a woman whose husband has died (18683)	
01 Relict, widow (18684)	widwe
02 Widower (18685)	wedener

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.

Curriculum vitae

Sander Stolk was born in Beverwijk on 5 March 1985. He obtained his VWO-Atheneum certificate from Kennemer College, Beverwijk, in 2003. In 2009, he received his M.Sc. degree in Computer Science at VU University, Amsterdam, where he specialized in Internet and Web technology. A subsequent study at Leiden University granted him a M.A. degree in Literary Studies, English Literature and Culture, in 2013. His M.A. thesis was awarded for being one of the most novel within the Humanities faculty.

Since 2014, Sander Stolk has worked at Semmtech, where he has employed linguistic principles to sustainably and effectively share information through ontologies and related Web technologies (i.e., Linked Data, Semantic Web) — as Software Developer (2014-15), Semantic Architect (2015-19), and, currently, Head of Innovation (2020-present).

Sander Stolk started his PhD research as an external student at Leiden University in 2016. He actively contributed to standardization efforts surrounding Linguistic Linked Data as a member of the Ontology-Lexica Community Group of W3C (2017-21); organized workshops on the subject of thesauri and applying Linguistic Linked Data to *A Thesaurus of Old English* (2019, 2020, 2021); co-edited a special issue of the journal *Amsterdamer Beiträge zur älteren Germanistik* 81.3-4 (2021); organized a session and round table discussion at the Leeds International Medieval Congress (2021); and co-founded and chaired the Digital Humanities Student Network (2017-22).

The majority of his publications and public speaking have either been on the efforts for his PhD research or, through his work at Semmtech, on the utilization of Linked Data for sharing information on the built environment.

List of publications

This section lists the author's publications that are on the subject of topical thesauri, Linguistic Linked Data, or both. These publications include conference papers, articles, specifications of Web vocabularies, and a book review. Those publications that appear as chapters within this dissertation are marked as such.

[2021]

Porck, M. H. and S. Stolk (eds.), *Exploring Early Medieval English Eloquence: A Digital Humanities Approach with A Thesaurus of Old English and Evoke*: Special issue of *Amsterdamer Beiträge zur älteren Germanistik* 81,3-4 (2021). https://brill.com/view/journals/abag/81/3-4/abag.81.issue-3-4.xml. Editorship of journal special issue

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), 318–58. doi: 10.1163/18756719-12340235. Article in journal Dissertation Chapter 7

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. Article in journal Dissertation Chapter 9

[2020]

Chiarcos, C., M. Ionov, J. de Does, K. Depuydt, A. F. Khan, S. Stolk, T. Declerck, J. P. McCrae, 'Modelling Frequency and Attestations for OntoLex-Lemon', Proceedings of Globalex 2020 Workshop on Linked Lexicography, 11-16 May 2020, pp. 1–9. https://www.aclweb.org/anthology/2020.globalex-1.1.pdf. Conference paper

[2019]

Stolk, S., '*A Thesaurus of Old English* as Linguistic Linked Data', Proceedings of eLex 2019, Sintra, 1–3 October 2019. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_13.pdf. Conference paper Dissertation Chapter 6

Stolk, S., '*lemon-tree*: Representing Topical Thesauri on the Semantic Web', Proceedings of Linked Data Knowledge 2019, Leipzig, 20–23 May 2019. doi: 10.4230/OASIcs.LDK.2019.16. Conference paper Dissertation Chapter 5

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

[2018]

Lemon-tree', ed. S. Stolk, 6 October 2018. https://w3id.org/lemon-tree. Web vocabulary

[2017]

Stolk, S., 'OntoLex and Onomasiological Ordering: Supporting Topical Thesauri', Proceedings of the LDK 2017 Workshops, NUI Galway, 18 June 2017, pp. 60–67. http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_3.pdf. Conference paper Dissertation Chapter 4

Stolk, S., '*Mapping English Metaphor through Time*, eds. W. Anderson, E. Bramwell, and C. Hough', *English Studies* 98.8 (2017), 1004–6. doi: 10.1080/0013838X.2017.1365556. Book review

List of source code

This section lists source code of software released by the author as part of this dissertation. These four releases are Web applications and data transformations that enable users to work with Linguistic Linked Data. The releases listed here have been made available under open licenses.

Evoke Web application
Creator: Sander Stolk
Version: 1.4.1
Released: 6 January 2022
License: GPL 3.0
Repository: https://github.com/ssstolk/evoke
Described in: Chapter 7
Mentioned in: Chapter 8, 9

Alignment tooling PDF glossaries and *TOE*-LLD Web application
Creator: Sander Stolk
Version: 1.0
Released: 12 October 2022
License: GPL 3.0
Repository: https://github.com/ssstolk/aligntool-pdfgloss
Described in: Chapter 7
Mentioned in: Chapter 8

***TOE* to *TOE*-LLD** Data transformation
Creator: Sander Stolk
Released: 5 November 2021
License: MIT
Repository: https://github.com/ssstolk/toe-lld
Described in: Chapter 6
Mentioned in: Chapter 7, 8, 9

'Old Frisian: Kinship' from MS Excel to LLD Data transformation
Creator: Sander Stolk
Released: 5 November 2021
License: MIT
Repository: https://github.com/ssstolk/oldfrisian-kinship
Described in: Chapter 7
Mentioned in: Chapter 8, 9

Samenvatting (Dutch summary)

Thesauri zijn waardevolle informatiebronnen binnen de wetenschap. Door hun onomasiologische ordening, van betekenis naar woorden welke die betekenis uitdrukken, zijn deze lexicografische bronnen waardevol voor taalkundig en cultureel onderzoek. De eenvoud waarmee zulk onderzoek uitgevoerd kan worden, is afhankelijk van de vorm waarin een thesaurus beschikbaar is gesteld. Om die reden verkent deze dissertatie een nieuwe publicatievorm voor thesauri, specifiek gericht op zulke werken die historische taalvarianten beslaan, waarbij wordt getoetst of de vorm additionele gebruikersfunctionaliteit kan faciliteren. De digitale vorm die wordt getoetst is Linguistic Linked Data, die mechanismen van het Web gebruikt om taalkundige gegevens aan elkaar te koppelen en daarmee een netwerk van informatie vormt dat kan worden bevraagd.

Het eerste deel van de dissertatie maakt zowel bestaande als gewenste karakteristieken inzichtelijk van thesauri van historische talen. Hoofdstuk 1 benoemt daartoe de elementaire bouwstenen waaruit de inhoud van thesauri bestaat. Dit overzicht berust zich op een analyse van beschikbare thesauri voor historische varianten van het Schots en het Engels. Een thesaurus, als zijnde een semantisch georganiseerd woordenboek, bestaat uit drie hoofdbouwstenen: (1) een hiërarchie van semantische begrippen; (2) woorden of frasen in een bepaalde betekenis, gepositioneerd binnen de eerdergenoemde hiërarchie; en, optioneel, (3) relaties van synonymie tussen woorden of frasen in een dergelijke betekenis. De daadwerkelijke informatie die is vastgelegd middels deze bouwstenen verschilt van thesaurus tot thesaurus. Zoals Hoofdstuk 2 aantoont, wensen onderzoekers de inhoud van thesauri te benutten als opstap voor het verkennen van gerelateerde materie die buiten het kader valt van het gebruik zoals initieel beoogd door de samenstellers van de thesaurus. De mogelijkheid om inhoud van thesauri te hergebruiken, uit te breiden en er naderhand onomasiologische analyses op te verrichten, zijn de noemenswaardigste nieuwe functionaliteiten die onderzoekers wensen en die vooralsnog ontbreken in gepubliceerde thesauri van de historische varianten van het Schots en het Engels.

Volgend op de analyse van de inhoud en functionaliteit gewenst voor thesauri van historische talen, beschouwen Hoofdstukken 3 tot en met 5 in welke digitale vorm deze lexicografische werken gepubliceerd zouden moeten worden op het Web. De voorgestelde vorm is gebaseerd op standaarden van het Semantic Web, met name SKOS en Lemon-OntoLex, en wordt recentelijk aangeduid als Linguistic Linked Data. De huidige specificaties van deze twee standaarden bleken niet afdoende om thesauri volledig vast te leggen; mijn werk rondom de ontwikkeling van het *lemon-tree* model was erop gericht om dit hiaat te vullen. Naast het aangeven op welke wijze de twee eerdergenoemde standaarden gecombineerd dienen te worden voor het vastleggen van thesauri, dekt *lemon-tree* twee verdere aspecten van thesaurusinhoud die van belang zijn voor dit doel: (1) niveaus die te onderscheiden zijn binnen de hiërarchieën van semantische concepten en (2) een lossere vorm van categorisering dan lexicalisatie. Deze toevoegingen zijn niet enkel relevant voor thesauri van historische talen, maar

tevens voor thesauri in het algemeen.

De laatste vier hoofdstukken evalueren het nut van de Linguistic Linked Data-vorm van thesauri door deze toe te passen op *A Thesaurus of Old English* (*TOE*), een thesaurus die de vroegmiddeleeuwse variant van het Engels beslaat. Het resultaat van de transformatie van *TOE* van het oorspronkelijke formaat naar Linguistic Linked Data (ook wel *TOE*-LLD genoemd), zoals besproken in Hoofdstuk 6, is beschikbaar gesteld in de webapplicatie Evoke. Deze applicatie is specifiek ontwikkeld voor thesauri als onderdeel van dit doctoraal onderzoek en wordt behandeld in Hoofdstuk 7. Evoke en *TOE*-LLD zijn gebruikt in verscheidene casussen binnen het onderzoeksproject 'Exploring Early Medieval English Eloquence' (EEMEE). Hoofdstuk 8 biedt een overzicht van deze casussen, inclusief reflectie op de waarde van de twee bronnen die erin centraal staan. Uit de evaluatie van de gekozen digitale vorm voor thesauri en de functionaliteit die Evoke biedt, komt naar voren dat de combinatie van de twee een krachtig middel kan zijn voor onderzoekers om een thesaurus met extra informatie te verrijken. Onderzoekers hebben aangetoond dat ze met deze middelen tot hun beschikking, vernieuwend onderzoek hebben kunnen doen naar, onder andere, historische ontwikkelingen binnen de lexicografie, stilistiek, diachrone taalontwikkelingen en verschillen tussen verwante historische talen.