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

Stolk, S.S.

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.

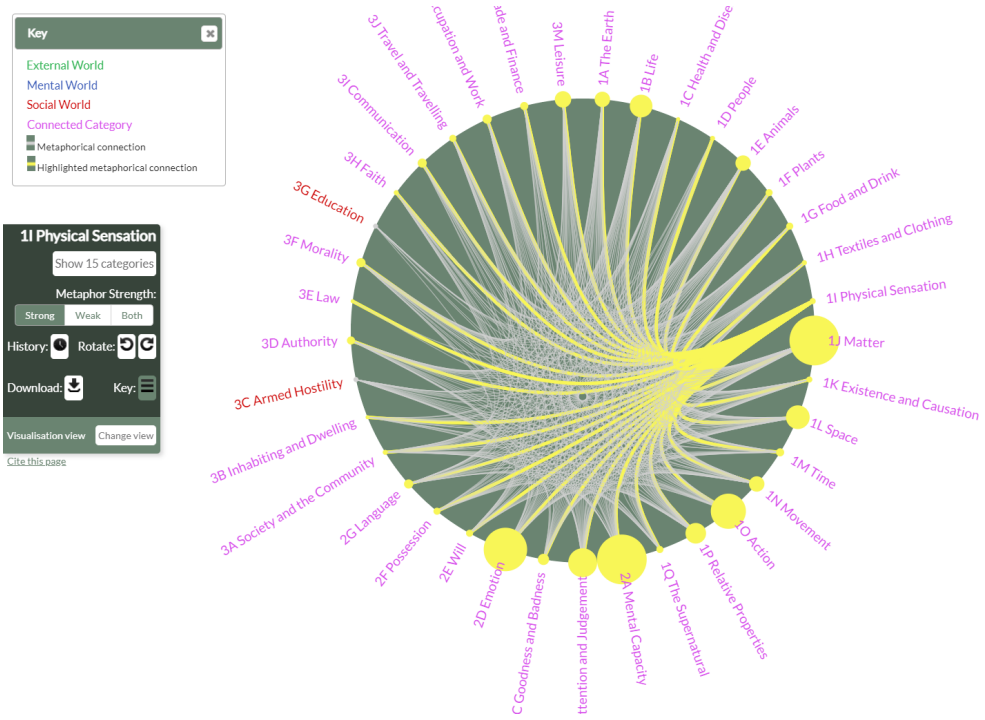

Figure 0.1.: Screenshot of the *Mapping Metaphor* website, which visualizes metaphorical connections between semantic fields.

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.

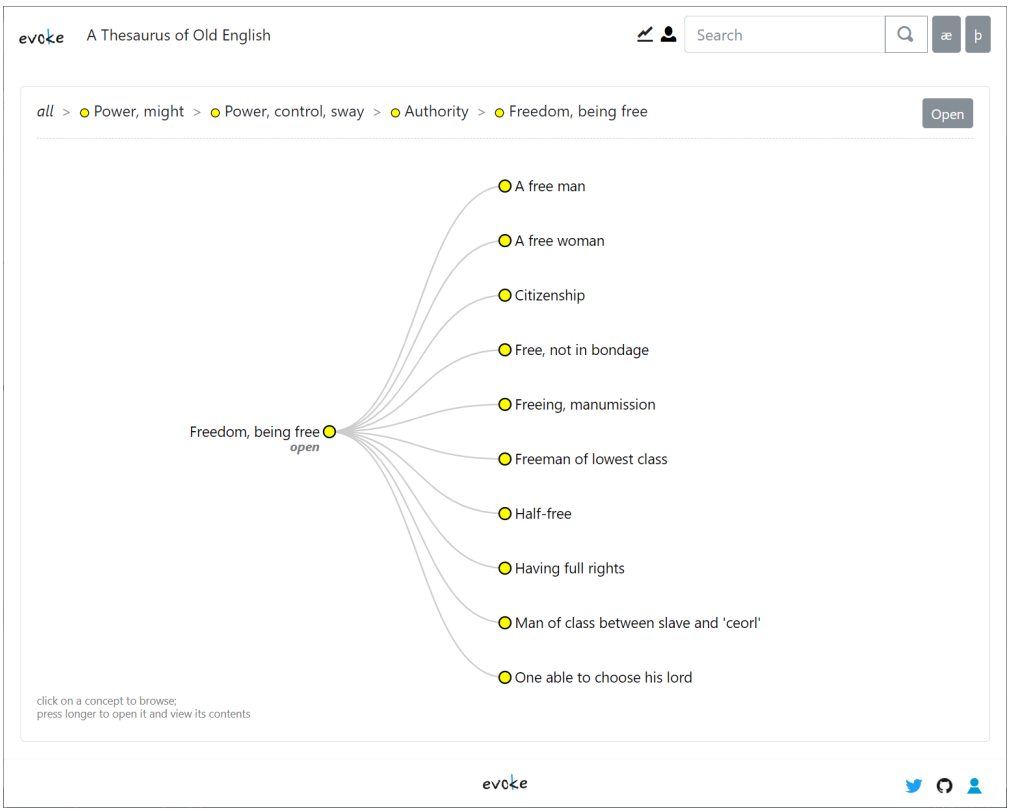

Figure 0.2.: Navigating the topical system of *TOE* in the web application Evoke.

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.