Lang Resources & Evaluation (2020) 54:615–643
https://doi.org/10.1007/s10579-019-09468-2

ORIGINAL PAPER

Digitization of data for a historical medical dictionary

Juhani Norri[1] · Marko Junkkari[1] · Timo Poranen[1]

Published online: 4 June 2019

Abstract What are known as specialized or specialist dictionaries are much more than lists of words and their definitions with occasional comments on things such as synonymy and homonymy. That is to say, a particular specialist term may be associated with many other concepts, including quotations, different senses, etymological categories, semantic categories, superordinate and subordinate terms in the terminological hierarchy, spelling variants, and references to background sources discussing the exact meaning and application of the term. The various concepts, in turn, form networks of mutual links, which makes the structure of the background concepts demanding to model when designing a database structure for this type of dictionary. The *Dictionary of medical vocabulary in English, 1375–1550* is a specialized historical dictionary that covers the vast medical lexicon of the centuries examined. It comprises over 12,000 terms, each of them associated with a host of background concepts. Compiling the dictionary took over 15 years. The process started with an analysis of hand-written manuscripts and early printed books from different sources and ended with the electronic dictionary described in the present paper. Over these years, the conceptual structure, database schema, and requirements for essential use cases were iteratively developed. In our paper, we introduce the conceptual structure and database schema modelled for implementing an electronic dictionary that involves different use cases such as term insertion and linking a term to related concepts. The achieved conceptual model, database

✉ Marko Junkkari
marko.junkkari@tuni.fi

Juhani Norri
juhani.norri@tuni.fi

Timo Poranen
timo.poranen@tuni.fi

[1] Faculty of Information Technology and Communication, Tampere University, FI-33014 Tampere, Finland

structure, and use cases provide a general framework for reference-oriented specialized dictionaries, including ones with a historical orientation.

Keywords Historical lexicography · Historical lexicology · Medical terminology · Database design · Digital dictionary

1 Introduction

Many modern dictionaries are published both as print and electronic versions, the quick searching features of the latter being particularly appreciated by users. There is, however, variation in the functionalities offered by individual works according to the target audience and its needs. Basic general-purpose dictionaries present simple descriptions of words and possible examples of their use. The database structure of these kinds of works is uncomplicated and straightforward to implement. Relationships between words and navigation via them, based for example on synonymy or hyponymy, are more advanced features. Their implementation requires a structural database solution where the mutual relationships between words are defined. In what are called specialist (Jackson 1988) or specialized (Hartmann and James 1998; Svensén 2009; Béjoint 2010; Becker 2016) dictionaries, notably special-subject dictionaries, a term is associated not only with other terms but also with different concepts behind the terms. For example, a term can be related to different semantic and etymological categories, it may carry different senses, a sense may be related to other senses, etc. Further, searching needs are not only focused on terms but also on the background concepts. This means that the corresponding database structure must be considerably richer and more complicated than in the context of basic general-purpose dictionaries. For designing the more advanced kind of database, the background concepts and their relationships must be analysed and modelled.

In the present paper, we design a conceptual data model for representing the concepts and relationships needed in a particular type of specialized dictionary. The conceptual model is transformed to a database schema, on top of which the actual application is implemented. The application supports advanced search features, data updating, linking instances of different concepts, and automated generation of printed or digital versions of the dictionary. The related application domain is medical vocabulary in late medieval England, the subject of long-standing linguistic research in the authors' home university. Over a period of 15 years of dictionary compilation, the requirements for the database and system have been focused on step by step. Although the conceptual model, database and system were developed for one specialized field, the conceptual model and database structure presented are not application domain specific, i.e. they do not contain application specific concepts.

The database that is the object of our interest was created for the *Dictionary of medical vocabulary in English, 1375–1550* (*DMV*; Norri 2016; see also Norri 2010). Unlike monumental works that aim at covering the entire lexicon of the centuries examined (see Sect. 2), *DMV* focuses on one area of vocabulary in late medieval England. Besides laying the foundation for the printed dictionary, the *DMV* database is also intended as a research tool and therefore contains information that is usually not given in a general-purpose dictionary, at least not in a systematic or explicit manner.

At the general level, we are interested in the requirements that a specialized dictionary, in particular a historical special-subject dictionary, presents on the digitization of data as regards conceptual modelling, database implementation, and application for managing the data modelled and collected. We introduce step by step how the conceptual model is designed, and what kind of database structure is derived from it. We also demonstrate the features of the application by describing the most typical use cases in searching and managing the dictionary data.

The rest of the article is organized as follows: in Sect. 2, the motivation for compiling a dictionary of medical vocabulary in late medieval England is discussed. Section 3 outlines the corpus of texts that were analysed for the dictionary. Section 4 provides an account of the structure of the database. The implementation of the system, with relevant screenshots, is described in Sect. 5, which is followed by a comparison of the *DMV* database with other dictionary formats in Sect. 6. The article ends with an evaluation and conclusion in Sect. 7. In the sections below, the term *specialized dictionary* is throughout used in the more specific sense of 'special-subject dictionary'.

2 Motivation for compiling *DMV*

The first published English–English vocabulary lists, appearing from the late fifteenth century onwards, were alphabetical collections of words from special areas such as law, theology, and medicine, accompanied by their explanations. Such lists were often appended to scholarly works for the benefit of the user (Schäfer 1989). Robert Cawdrey's *A table alphabetical*, published in 1604, is the first book devoted to explaining difficult words in general use in English (Osselton 2009). Since those days, there has been a steady flow of English–English glossaries and dictionaries, motivated by the increase in literacy and the spread of the English language to all continents of the world. Many of the dictionaries of English that are published today are general-purpose dictionaries aiming at a comprehensive listing of the vocabulary, another major type being learner's dictionaries intended for non-native learners of English, with a focus on the more frequently used words and phrases. For scholarship on the English language, in particular lexical studies, the availability of three major historical dictionaries is crucial. The period C.E. 600–1150 is covered by the *Dictionary of Old English* (*DOE*; eds. Cameron et al. 1986), of which the letters A–I have now been completed. The *Middle English dictionary* (*MED*; eds. Kurath et al. 1952–2001), the result of a fifty-year project, focuses on the words used between C.E. 1100 and C.E. 1500. For subsequent centuries, including the vocabulary of present-day English, *The Oxford English dictionary* (*OED*; eds. Simpson et al. 2000), whose third edition is currently under preparation, is widely regarded as an unsurpassed authority. All three dictionaries are now available as searchable electronic versions, and the challenges posed by their computerization have been discussed in a variety of articles and books (e.g. Gray 1986; Healey 1985; Venezky 1988; Logan 1991; Elliott and Williams 2006; Weiner 2009; Gilliver 2016).

Owing to the important developments in dictionary-making represented by *DOE*, *MED*, *OED*, and many other lexicographic works, our knowledge about the history of English vocabulary is increasing all the time. A good deal of further work remains, however, to be done in recording the words employed in various areas of scientific study in the course of the centuries. Becker (2016) notes that as concerns both dictionary-making and dictionary research, historical specialized lexicography is an area that has received scant attention. Becker states that there is a particular need for dictionaries devoted to the origins and histories of scientific terms. In studies of dictionaries, it has been observed that lexicographers are faced with multifarious problems when dealing with scientific vocabulary, among them questions of inclusion and exclusion, semantic shifts, the degree of detail in the definitions, the wording of the definitions, the selection and role of the corpus, the status of individual terms, the use of subject labels, and the inclusion of symbols and formulae (Curzan 2000; Hoare and Salmon 2000; Becker 2016; Gilliver 2016). McConchie (1997) examined the vocabularies of thirteen medical books published between 1547 and 1612, some of them listed in the *OED* bibliography. He discovered 3985 items of new data for *OED*, including 2558 occurrences of a word antedating the dictionary citations, 1089 unrecorded words, 246 new senses, and 92 occurrences of a word postdating the last dictionary citation. When material from pre-1547 medical works, whether manuscripts or books, is subjected to lexical analysis, the results are likely to include a multitude of such items of new information. There has been a growing interest in Old and Middle English medical and scientific writings over the last few decades, a major milestone being the publication of the electronic Voigts and Kurtz database (2014–) of such writings. For the first time, scholars now have a comprehensive and systematic inventory of the medical and scientific works written in English before 1500. The research tool makes it possible to locate material of potential interest for lexicographic and lexicological studies, found in various libraries and private collections around the world.

There have been only a few dictionaries or glossaries focusing specifically on the history of English medical vocabulary. Members of the medical profession have produced a number of books outlining the history of terms still in use. Skinner (1961) and Haubrich (2003) have an alphabetical arrangement, the former providing information about the origins of c. 4000 terms, the latter describing the etymologies of c. 3300 terms. Two other doctors, Roberts (1980) and Dirckx (1983), divided the words and phrases that they discuss into sub-groups according to the source of the term. Language scholars have mostly dealt with the histories of a limited number of medical words in their publications, some examples being Foster's (1970) and Woledge's (1970) observations on the French provenance of Middle English *jowe* 'jaw', Cameron's (1988) discussion of the enigmatic maladies *þeor* and *þeoradl* in Old English manuscripts, Thompson's (1992) findings on the use of the word *cancer* in Old English, Voigts and Hudson's (1992) survey of the anesthetic called *dwale* in

Middle English, and Norri's (2017a) proposed identification of the mysterious body part *mould* in Middle English.

DMV is the first extensive dictionary devoted to late Middle English and pre-1550 Early Modern English medical vocabulary. The dictionary provides an inventory of contemporary names of body parts, sicknesses, instruments, and medicinal preparations. It took some 15 years to complete, and the process presented numerous challenges for the devisers of the database. The motivation for embarking upon the project was twofold. Firstly, the years 1375–1550 were highly significant for the development of an English medical vocabulary, as it was in the late fourteenth century that English began to emerge from the shadow of Latin and French as the language of medical writing in England (Voigts 1995). Many of the words used by the writers and translators did not, however, gain any wider currency and have not been recorded in any post-1550 texts. The 175 years were characterized by terminological instability, with plenty of lexical entrances and exits (Norri 2004). A second reason for examining medical vocabulary of the period 1375–1550 was its uneven coverage in existing historical dictionaries of English. For practical reasons, *MED* and *OED* mostly analyse the vocabularies of printed works, including editions, and cannot devote similar attention to texts that only exist as manuscripts. The latter constitute a largely untapped resource for lexicological and lexicographic studies. The first printing-press in England was set up in 1476, which means that most Middle English medical writings were handwritten manuscripts. Some of them were later published as books (McConchie 1988; Voigts 1989) and are included in the corpora for *MED* and *OED*. It has been observed that the compilers of the two dictionaries were somewhat inconsistent in their treatment of scientific and medical words, especially as concerns inclusion or exclusion of special vocabulary of Latin or French origin (Norri 1992; McConchie 1997; Landau 2001).

The target audience of *DMV* are lexicologists, lexicographers, editors of medical manuscripts, and historians of medicine. For anyone interested in studying Middle English and Early Modern English medical works, the terminology employed by contemporary writers presents a major challenge. Many of the words and phrases have disappeared without a trace in the course of the centuries. For terms that are still used in modern medicine, the meaning is often different in the medieval text. The database underlying *DMV* is searchable and enables users to collect data on a variety of aspects relating to medical words and phrases in late medieval England. Examination of the general characteristics of medical vocabulary between 1375 and 1550 would be a painstaking effort using the circa 1300-page printed volume as the main source.

3 Dictionary corpus

The corpus gathered for the dictionary comprises altogether 11,397 pages from late medieval manuscripts and printed books. In collecting the corpus, the printed manuscript catalogues of various libraries were first consulted. For locating medical books published during the period examined, the catalogue of early printed books by Pollard and Redgrave (1986–1991) was the main source. The texts that seemed most promising were then checked personally in the libraries of Oxford (Bodleian Library), Cambridge (Cambridge University Library, the libraries of Emmanuel College, Gonville and Caius College, Jesus College, Magdalene College, Peterhouse, and Trinity College), and London (the British Library, the Wellcome Library). Treatises that proved to be short, or of a fragmentary nature, were not included, but both acephalous and atelous texts were admitted in those cases where an important medical work only survived in a unique incomplete version. Medical writings in which languages other than English dominated were excluded from the corpus.

One of the guiding principles in choosing the texts to be examined was inclusion of different types of medical treatises written in medieval England. Voigts (1982) introduced a now widely followed classification of Middle English medical writings based on the origin of the text and the tradition behind it. Voigts makes a distinction between remedybooks and academic medical texts. Remedybooks form the older and larger group. They consist mainly of medicinal recipes, whose contents can be divided into six components: the purpose (e.g. “For tooth ache”), the requisite ingredients and equipment, the preparation of the medicine, its administration, the rationale (the reason for trusting the remedy prescribed), and finally incidental data such as anecdotes of successful cures (Stannard 1982). In addition to recipes, many remedybooks contain other kinds of material relating to the cure of illnesses and the maintenance of health.The reader may come across passages relating to for example blood-letting, diet, and the magical properties of plants and objects. The origins of particular remedybooks are difficult to trace, as the compilers freely excerpted their material from sources that they mostly do not identify. Repeated copying by scribes whose knowledge of medical matters varied sometimes introduced mistakes into the text, including garbled or corrupt versions of names of body parts, sicknesses, and medicinal ingredients (Keiser 1978; Benskin 1985; Norri 1988a, b).

The second category in Voigts’s classification is Middle English versions of academic medical texts, which, besides healing methods, also discuss the causes of disorders of body or mind, their symptoms, and prognosis. These treatises derive from the university tradition, comprising works written by classical and medieval medical authorities. Although different copies of the same work may show instances of simplification or condensation, academic medical texts underwent far less revision than remedybooks (Voigts 1984). Especially common in this group are surgical manuals, which are mainly translations from Latin. Earlier studies of Middle English and Early Modern English medical vocabulary have shown that the terminology employed in surgical manuals differs quite significantly from that found in other types of academic medical writing. Of the altogether 2364 names of sicknesses and body parts that were collected for two earlier glossaries (Norri 1992, 1998), 810 (34%) occurred only in surgical manuscripts and books. Terms that were exclusive to non-surgical academic works numbered 507 (21%). The medical vocabulary of remedybooks tends to be somewhat rudimentary, and there were only 136 terms (6%) restricted to those works. There is thus some justification for making a threefold rather than a twofold classification of source material in studies of medical vocabulary in late medieval England.

The corpus collected for *DMV* comprises medical texts from all three categories discussed above. In terms of manuscript and book pages, surgical manuals make up the single largest group, with some 5000 pages taken from such works. The other types of academic treatise, which include general compendia on medicine as well as tracts devoted to specific subjects (e.g. uroscopy, fevers, gynecology, obstetrics, distillation of medicines), total approximately 4000 pages. Remedybooks are the smallest section of the corpus, running to circa 2500 pages. As noted above, the medical terminology found in remedybooks mostly consists of common words that also occur in surgical and other types of academic treatise. Some of the recipes have in fact been excerpted from academic works, but the technical terminology of the latter has been simplified or omitted in the process (Getz 1982). Owing to the basic nature of the vocabulary employed, remedybooks, unlike academic works, contain few explicit comments on the meanings of terms. In lexical studies, there is thus a good case for concentrating on academic and surgical works rather than remedybooks.

The corpus contains unedited manuscripts, editions of Middle English medical works, and early printed books. One of the aims when collecting the material was to include a large number of unedited manuscripts. It was expected that previously untapped manuscript material would bring to light many words and phrases as yet unrecorded in the existing historical dictionaries of English. Of the 72 different medical treatises in the corpus, 38 are found in manuscripts, 18 in editions, and 16 in early printed books.

The gathering of the corpus began in the late 1980s, and the last texts were selected for it at the beginning of the present millennium. In the reading programme, the greatest challenge was posed by handwritten manuscripts, ordered from the various libraries as microfilms. In the treatment of the data, methods of optical character recognition could not be used, because they have little precision for handwritten text (Sonkusare and Sahu 2016). The methods of ORC have developed in recent times but there is still room for further improvement.

4 Description of the structure of the *DMV* database

The ER (Entity–Relationship) model is an established conceptual modelling method for representing concepts, their mutual relationships, and the properties of concepts and relationships. A well-designed ER schema provides a sound framework for the structure of the database of the underlying application domain. The basic primitives of the ER model are the entity type (also called concept), relationship, and attribute. A concept involves a set of entities that are instances of the concept. A relationship is associated with one or more concepts, and attributes are attached to concepts or relationships. A relationship also involves cardinality constraints that determine whether one or more instances of concepts participate in the relationship.

In an ER schema, a concept is represented by a rectangle, an attribute by an oval, and a relationship by a rhomb. In Fig. 1, a fragment of an ER schema is given. In it, three concepts and their mutual relationships are represented. The relationship between TERM and TERM_SOURCE involves *n*:1 cardinality, which means that a

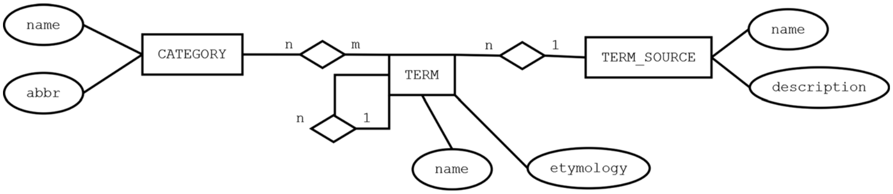


Fig. 1 Sample ER schema

term is associated with one source but a source may be associated with several terms. The *n*:*m* cardinality between TERM and CATEGORY means that a term may belong to several categories and a category may contain several terms. There is also a relationship within TERM. This kind of relationship is called a recursive relationship and it can be used to express relationships of sub- and super-concepts.

We use the ER model to represent the concepts and their relationships pertaining to the present project. First, ER schema fragments are given to illustrate the data structures needed in different use cases. Then, the fragments are collected into one ER schema, in terms of which the structure of the database is formed. Below, the attributes are not given in the visual representation so as to keep the illustrations compact. TERM, for example, contains almost twenty attributes and their visual representation would make the schema large and complex to follow. The associated attributes are given in connection with the database schema in "Appendix".

4.1 Quotations and related concepts

The texts in the corpus were subjected to a systematic reading programme, during which any occurrences of words that were potential candidates for inclusion in the dictionary were noted down. Those items that on closer inspection were judged to be useful additions to the lexical inventory (e.g. rare or previously unrecorded words) were stored in the database. The first stage in the storing process was the creation of a QUOTE containing the word and some of the passage surrounding it in the medieval work. The length of the quotation varied depending on how much of the context was illuminating for the Middle English or Early Modern English use of the word. The following three quotations were stored in the database because they provide useful definitions of the common medical terms *shingles*, *polyp*, and *fistula*, respectively (cf. the discussion of "defining quotations" in Lewis 2007):

> For þe schyngyllys, here ys a pryncypal medysyn. The malady in maner ys lyke wylde fyyr and yt wyll sprynge owte and ryn abowte a man.

> Polippis..is fleisch þat growiþ wiþinne þe nose.

> Fistula is an hollowe sore, and it is so called bycause it hath an holownes lyke a pype. For the same cause the Grecians haue named it syrynges.

The TEXT from which the quotation comes was identified, together with the TEXT_TYPE (manuscript, edition, printed book). Thus, for example, the first

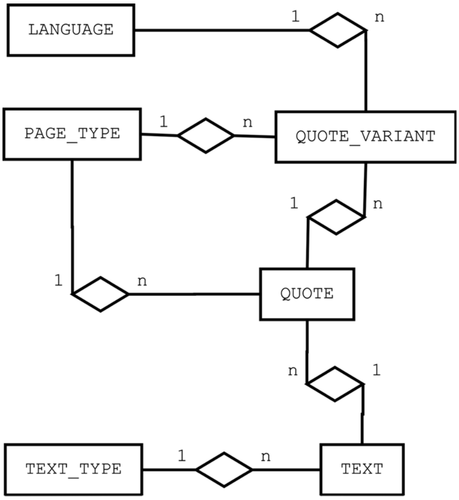


Fig. 2 Concepts involved in storing quotations

quotation above is taken from a remedybook found in an unedited manuscript (abbreviation **RemAshm.1389* in *DMV*), the second from an edition of the surgical work by the Milanese surgeon Lanfrank (*Lanfrank(1)* in *DMV*), and the third from the glossary of the printed book containing the surgical manual by the Genoese Joannes de Vigo (*VigoGloss* in *DMV*).

For locating the passage in the original work, information about the page or leaf in question is essential. The unedited manuscripts in the corpus were either paginated (each page numbered) or foliated (each leaf numbered). As for edited manuscripts, the most convenient guide to the relevant passage is the page number of the edition, followed by the line where the quotation begins. Early printed books were assembled from gatherings of leaves, each gathering signed by a letter of the Latin alphabet, with leaves subsequent to the first carrying a numerical suffix, e.g. A2, B4 (see Gaskell 1995). The concept PAGE_TYPE was included in the database structure for covering the various alternatives of numbering and alphabetization, which for the above quotations would be p. 230 (**RemAshm.1389*), p. 19, l. 16 (*Lanfrank(1)*), and signature Zz6rb, that is, the second column of the recto of the leaf signed Zz6 (*VigoGloss*).

For each quotation, the word or words of interest therein needed to be identified. English spelling was not yet standardized, and the same word could be spelt in a variety of ways (e.g. *chyngelles*, *schyngelys*, *schyngyllys*, *sengles*, *shyngles*, etc.), which explains why the relevant concept was named QUOTE_VARIANT. Some of the terms for body parts, sicknesses, medicines, and instruments are said to be used in languages other than English by the medieval authors. Formulae like “in Greke it is named Cardia” and “whyche the Arabians call Nucha” occur in both Middle English and Early Modern English medical treatises (Norri 1992, 1998; Pahta 2011). Such words were also collected from the corpus for research purposes, although the dictionary itself lists only words that appear in an English context, without being labelled foreign. The information in the database specifies the language status of the word in each quotation (LANGUAGE). Both *fistula* and *syrynges* in the *VigoGloss* quotation above were singled out as relevant terms, the language specification for the former being English (ENG), for the latter Greek (GRE).

The fragment of the ER schema in Fig. 2 summarizes the concepts discussed so far.

4.2 Variants, terms and senses

Some of the variant spellings of a word occurred frequently in the quotations, others only once. All the different variants collected from the quotations formed a concept of their own (VARIANT). For the word *heart*, for example, the list of variant spellings is the following: *ert*, *hart*, *harte*, *heart*, *hearte*, *hert*, *herte*, *hertte*, *horte*. In historical dictionaries, it is customary to indicate separately whether a particular spelling variant is a singular or plural form. Owing to limitations of space, printed dictionaries do not usually give all the regular plural forms ending in *-s* if there are many of them and the list of singular forms already contains the same spelling variant without the plural *-s* (see e.g. Lewis 2007). In the *DMV* database, the concept VARIANT_USAGE relates to such issues. Whether a particular variant is a singular (e.g. *knee*) or plural (e.g. *knees*) form, or possibly either (like *axces* 'attack(s) of fever'), is specified. The need to list the variant separately in the printed dictionary is stated, as is the need to provide a cross-reference from the variant to the dictionary headword (TERM) in cases where the two are quite different in form (e.g. from spellings like *eighe* or *iye* to *eye*).

In medieval medical vocabulary, polysemy was rife. It is therefore important to specify the meaning in which the word is used in a particular quotation. For this purpose, pairs of headwords and senses were formed (TERM_SENSE), combining information from the concepts TERM and SENSE (of the former, more below). Each variant occurring in the quotations, except for the words said to belong to some foreign language, was linked to the relevant TERM_SENSE pair. The following nine pairs, for example, were created to cover the different meanings of the word *bag* that emerged in the course of the lexical analysis:

1. BAG_bag, pouch, small sack
2. BAG_bag containing medicinal mixture applied externally
3. BAG_pouch for mastication, filled with herbs and spices
4. BAG_bag-like support of bandages, used in the treatment of e.g. penile ulcers
5. BAG_stomach, sac between esophagus and duodenum
6. BAG_gall bladder
7. BAG_fetal membranes, placenta and umbilical cord; extruded from uterus after birth
8. BAG_cyst, abnormal sac within body, containing liquid or semi-solid material
9. BAG_mass of morbid matter

In the setting up of the TERM_SENSE pairs, questions of number proved to be a challenge. Occasionally, only the singular or the plural form of the word was used in

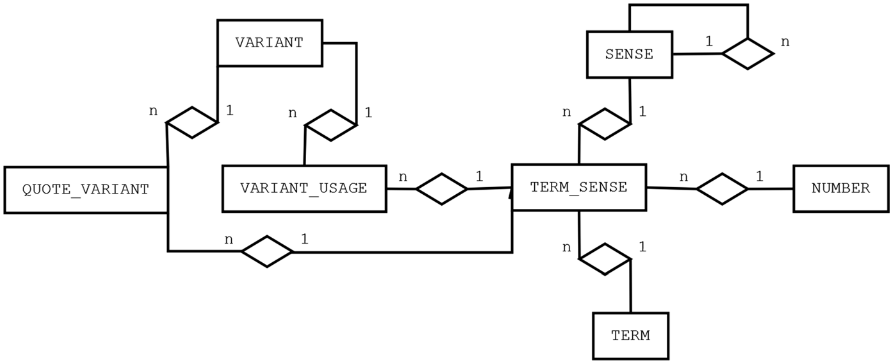


Fig. 3 Concepts involved in linking variants to senses

a particular meaning. In some instances, singular and plural forms were used interchangeably, without a change of meaning: both singular *gut* and plural *guts* meant 'intestine', and *inward* and *inwards* both denoted the internal organs of the body (*DMV* s.vv. *gut*, *inward*). For indicating relations between individual meanings and grammatical number, the concept NUMBER was included in the schema.

Figure 3 provides a summary of the concepts that are central in linking the variant spellings occurring in the texts to the specific senses that the word may carry.

4.3 Terms and lexical research

The different spelling variants of basically the same word in the corpus were placed under a particular headword (TERM). The form chosen as the headword was the one closest to the corresponding modern term (if any). The LEXICAL FIELD of the headword could be one or more of the four areas of vocabulary studied, i.e. body parts, sicknesses, instruments, and medicines. Because of polysemy, multiple placement was common. As appears from above, one of the most polysemous words in late medieval medical writings was *bag*, used of body parts (senses 5, 6, 7), sicknesses (8, 9), instruments (1, 4), and medicines (2, 3) alike.

The *DMV* database further provides the etymology of each headword, the concept CATEGORY summarizing the origin. The nineteen categories include groups such as French loanwords, Latin loanwords, English suffix formations, and English prefix formations, to cite but a few examples. The concept CATEGORY GROUP provides a more general classification of the nineteen categories into six larger groups, among them simple (non-affixed) terms and derivatives of foreign origin, which includes loanwords from e.g. French and Latin. Another category group consists of derivatives of English origin, which comprise English prefix and suffix formations.

If the medieval term has been discussed by earlier scholars in their publications, references are given to such articles and books (BOOK_REFERENCE). Those

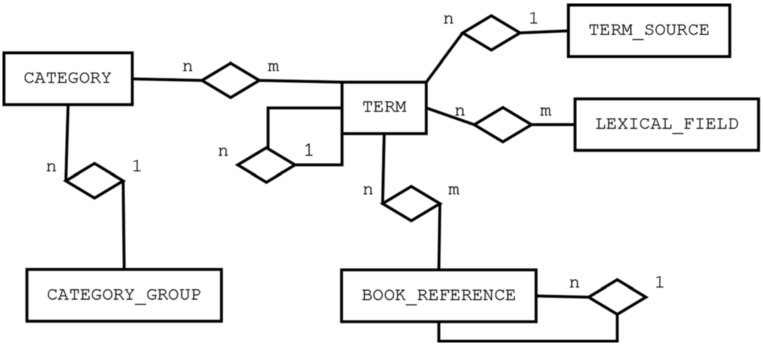


Fig. 4 Concepts involved in classifying the term for lexical research

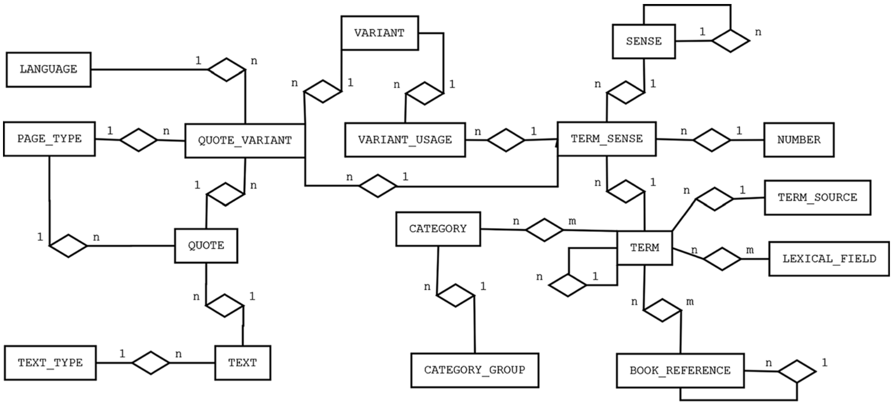


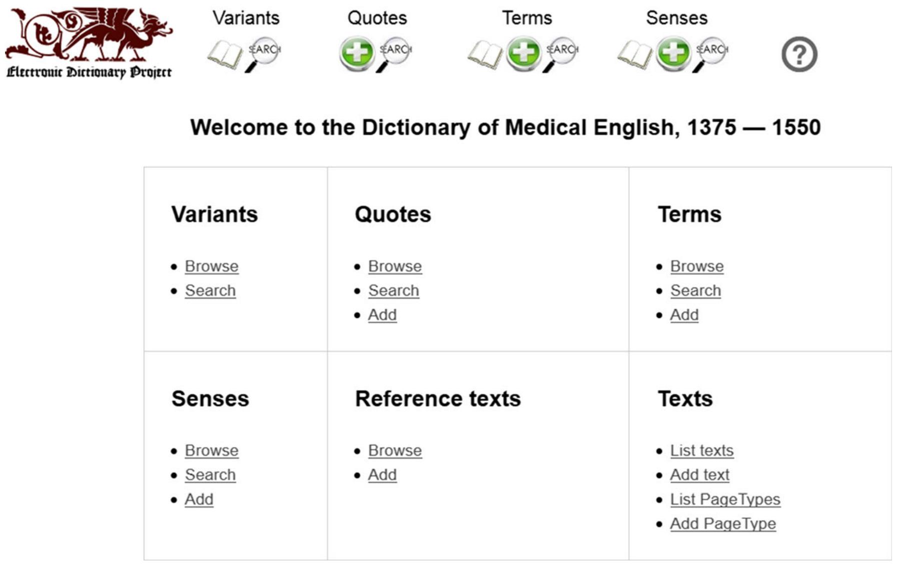
Fig. 5 Full set of concepts and their relationships

terms recorded by *OED* or *MED* lexicographers are indicated, as are words whose first attested occurrence comes from the *DMV* corpus (TERM_SOURCE).

Figure 4 shows the concepts that are used either to categorize the terms for purposes of lexical research or to indicate earlier studies on particular terms.

4.4 Summary of the concepts central to *DMV*

Now that the different concepts and their relationships have been discussed in the light of three fragments of the ER schema (Figs. 2, 3, 4), it is appropriate to present a diagram of the entire ER schema, found in Fig. 5. The fragments are integrated based on the common concepts found in them. In other words, QUOTE_VARIANT connects the schemata in Figs. 2 and 3, whereas TERM connects the schemata in Figs. 3 and 4.



Welcome to the Dictionary of Medical English, 1375 — 1550

Variants	Quotes	Terms
• Browse • Search	• Browse • Search • Add	• Browse • Search • Add
Senses	**Reference texts**	**Texts**
• Browse • Search • Add	• Browse • Add	• List texts • Add text • List PageTypes • Add PageType

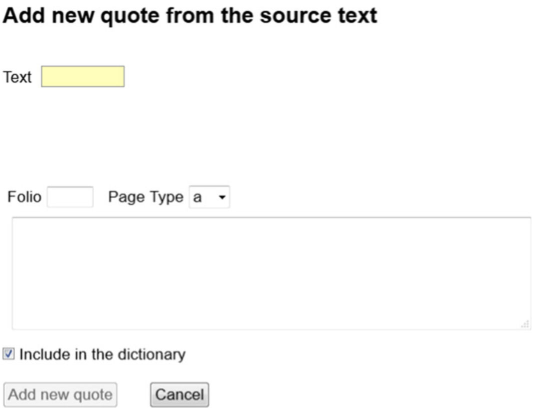
Fig. 6 Main page of the dictionary

"Appendix" (Fig. 13) presents the database schema formed based on the ER schema in Fig. 5. The database consists of tables that correspond to the concepts and relationships of the ER-model. A table involves attributes (properties) having different roles. All the tables involve the id attribute (primary key) that identifies the instances (rows) of the table. An arrow from a foreign key (e.g. term_id) to a primary key (id) describes a reference between two tables. The rest of the attributes contain actual data. For example, the 'categories' table has an id attribute, a reference attribute (category_group_id), and two data attributes (name and abbreviation). We do not introduce the data attributes in detail here, but next we will illustrate their content by the application of *DMV*.

5 Use cases of *DMV*

During the years 2000–2006, the Paradox database created by the first author was used to collect and maintain dictionary data. The database structure became complicated as new items of information were incorporated, and redesign was necessary to improve the usability and performance of the database. At the same time, there arose a need to use a modern graphical user interface and enable the generation of a printed version of the dictionary.

Three student project teams (Poranen 2007; Kajaste and Poranen 2008; Poranen 2009) developed the *DMV* application during the years 2006–2009 under the supervision of the authors of the present paper. The first team converted the Paradox database to the PostgreSQL database management system (PostgreSQL 2018) and also implemented the first version of the graphical user interface using PHP



Page Type a

☑ Include in the dictionary

Fig. 7 A screenshot of the user interface for storing quotations

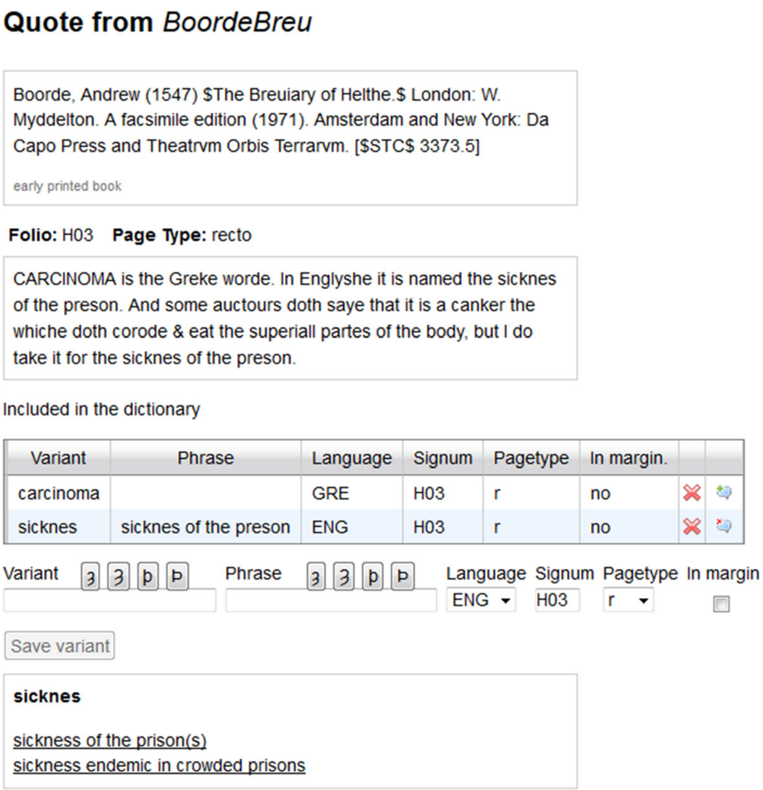
programming language. The second team refactored implementation to use the CakePHP framework (2018) and redesigned the user interface. The third team enhanced the functionalities of the system and made possible the generation of the printed version of the dictionary from selected parts of the database. The database structure was gradually moulded into its current form during these projects.

In what follows, the database is described from the point of view of those who use it for storing, browsing, or searching information. The different use cases set requirements on the functionalities of the system and on the kinds of information that need to be included therein. The main page of the database, shown in Fig. 6, presents an overview of the purposes for which the application can be used. We will discuss the most central functionalities in an order that roughly corresponds to the presentation of the concepts in Sect. 4, starting from the storing of quotations.

5.1 Storing quotations and variants

The screenshot in Fig. 7 shows the view for storing quotations from the corpus texts. The information that is stored comprises the following items: (1) an eight-character shorthand reference to the text in question (e.g. 1547BOOR for Andrew Boorde's *The breuiary of helthe*, published in 1547), (2) folio number, page number, or signature letter and number (cf. Section 4.1; called just "Folio" here), (3) the various special markings that foliation, pagination, or signatures require (e.g. *r[ecto]*, *v[erso]*, *[column] a*, *[column] b*; "Page type"), (4) the quotation itself, and (5) whether the quotation is intended for inclusion in the printed dictionary or stored for some other reason (e.g. because it contains a previously unattested spelling variant of a term).

When the user has completed the above information and clicks the "Add new quote" button, the screen that follows looks like the one in Fig. 8, except for the fact that the table of variants is still empty and the last box containing information about meaning is not yet there. The user inserts the variant(s) of interest into the table, specifying whether they are used as English words (ENG) or said to be foreign, e.g. Greek (GRE). As in Fig. 8, the medical term is often a longer phrase (e.g. *sickness*



Quote from *BoordeBreu*

Boorde, Andrew (1547) $The Breuiary of Helthe.$ London: W. Myddelton. A facsimile edition (1971). Amsterdam and New York: Da Capo Press and Theatrvm Orbis Terrarvm. [STC 3373.5]

early printed book

Folio: H03 **Page Type:** recto

CARCINOMA is the Greke worde. In Englyshe it is named the sicknes of the preson. And some auctours doth saye that it is a canker the whiche doth corode & eat the superiall partes of the body, but I do take it for the sicknes of the preson.

Included in the dictionary

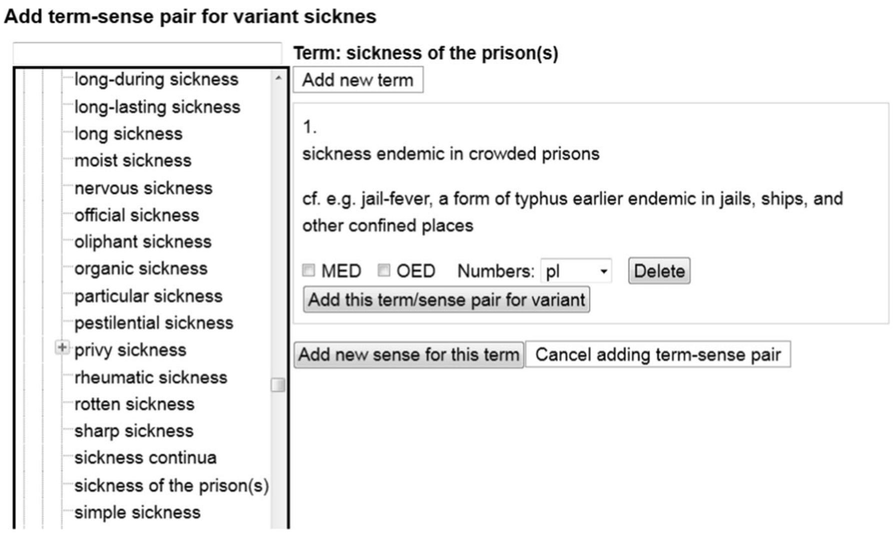
Variant	Phrase	Language	Signum	Pagetype	In margin.		
carcinoma		GRE	H03	r	no	✗	
sicknes	sicknes of the preson	ENG	H03	r	no	✗	

Language: ENG Signum: H03 Pagetype: r In margin: □

sicknes

sickness of the prison(s)
sickness endemic in crowded prisons

Fig. 8 A screenshot of the user interface for storing variants and information on them



Add term-sense pair for variant sicknes

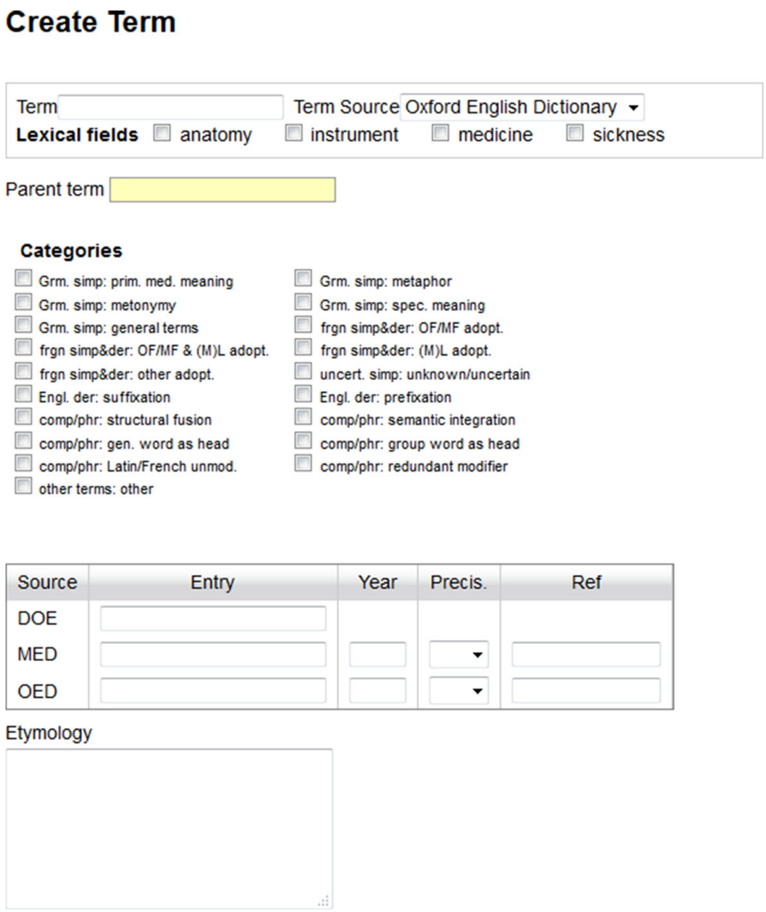
Term: sickness of the prison(s)

1.
sickness endemic in crowded prisons

cf. e.g. jail-fever, a form of typhus earlier endemic in jails, ships, and other confined places

□ MED □ OED Numbers: pl

Fig. 9 A screenshot of the user interface for adding a term-sense pair to a variant



Create Term

Term ______ Term Source Oxford English Dictionary ▾

Lexical fields □ anatomy □ instrument □ medicine □ sickness

Parent term ______

Categories

- □ Grm. simp: prim. med. meaning
- □ Grm. simp: metonymy
- □ Grm. simp: general terms
- □ frgn simp&der: OF/MF & (M)L adopt.
- □ frgn simp&der: other adopt.
- □ Engl. der: suffixation
- □ comp/phr: structural fusion
- □ comp/phr: gen. word as head
- □ comp/phr: Latin/French unmod.
- □ other terms: other
- □ Grm. simp: metaphor
- □ Grm. simp: spec. meaning
- □ frgn simp&der: OF/MF adopt.
- □ frgn simp&der: (M)L adopt.
- □ uncert. simp: unknown/uncertain
- □ Engl. der: prefixation
- □ comp/phr: semantic integration
- □ comp/phr: group word as head
- □ comp/phr: redundant modifier

Source	Entry	Year	Precis.	Ref
DOE				
MED			▾	
OED			▾	

Etymology

Fig. 10 A screenshot of the user interface for creating a term

of the prison) of which the variant is the head or main element. The medical term sometimes occurs in the margin of the manuscript or book, as a pointer to the adjacent passage. Such location outside the main text is also indicated in the table.

Each English-language variant needs to be placed under a particular dictionary headword and the sense that it carries in the quotation needs to be stated. For doing this, each line of variants ends with a thought bubble, whose clicking takes one to the comprehensive list of term-sense pairs. Figure 9 displays an extract from that list, showing how the variant *sicknes* and the phrase *sicknes of the preson* were linked to the term-sense pair *sickness of the prison(s)* ‘sickness endemic in crowded prisons; cf. e.g. jail-fever, a form of typhus earlier endemic in jails, ships, and other confined places’. If no suitable term or sense can be found, they will need to be created in the manner explained below. The button “Add new sense for this term” applies to cases where the term has already been created but has not yet been linked to a sense that would fit the quotation being worked upon.

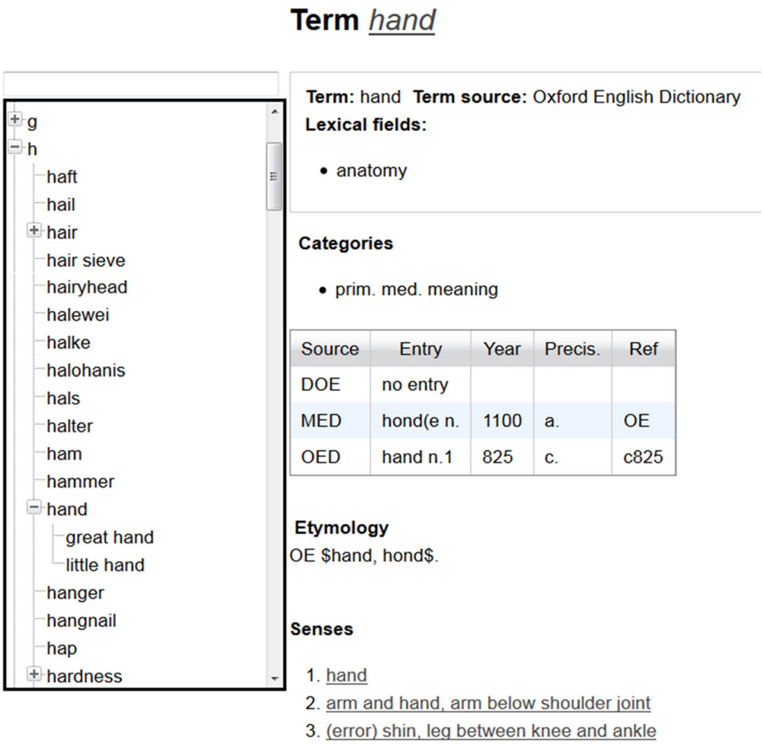
5.2 Creating new terms and senses

How a new term or dictionary headword is created appears from the screenshot in Fig. 10. The first item of information to be stored is the exact form of the headword under which the different spelling variants are placed ("Term"). The "Term source" (cf. Sect. 4.3) can be one of four possibilities: *Oxford English dictionary* (word also found in *OED*), *Middle English dictionary* (word also found in *MED*, but not in *OED*), Early Modern English (1500–1550) texts in the corpus, Middle English (1375–1500) texts in the corpus. A distinction is thus made between terms already listed in the major historical dictionaries covering the period 1375–1550 and terms whose first attested occurrence comes from the corpus texts. Further items of information that are provided include the lexical field(s) to which the term belongs, its etymological category, and a more detailed description of the etymology. A possible "Parent term" is indicated, that is, a superordinate term directly above the term being created (e.g. *sickness* in the case of *sickness of the prison(s)*). Any occurrences in *DOE*, *MED*, and *OED* are noted down, together with the dating of the term in *MED* and *OED* (in *DOE* no dates or years are usually cited). Information about the precision or exactness of that dating is also given, including the presence of a question mark or an abbreviation such as *c*. for *circa*.

For adding a new sense to the database, the screen is simpler and involves three types of information. In a specialized dictionary, the description of the meaning of a word or phrase often needs to be fuller than in a general-purpose dictionary. In the latter kind of work, it may be enough to state that *save* was a medicinal paste used internally in the treatment of wounds and injuries, but medical historians and editors of medieval medical treatises would surely also like to know the ingredients and the method of preparing the medicine. Such lengthy or encyclopedic descriptions would, however, be unwieldy in a comprehensive list summarizing the senses that medical words had during the period investigated. This is why the new senses that are stored in the database are divided into two main parts. One is the core meaning, or the most essential definition ("Sense"), the other being an elaboration of that core meaning or addition of information of an encyclopedic nature ("Description"). When senses are browsed, only the former are shown on the list. The user adding a new sense further needs to specify whether it is a top-level (root) sense or has a parent sense above it in the semantic classification (cf. the previous notion of "parent term"). Thus, for example, 'lower jawbone', 'jawbone', 'bone', and 'part of the body' form a semantic chain where each sense has the following as its parent until 'part of the body' is reached.

5.3 Carrying out searches

It is possible to carry out searches on variants, quotations, terms, and senses stored in the database. The searches for variants, quotations, and senses are straightforward in that they involve typing a spelling variant, key word, or several key words in the search box. The search for terms offers a wider variety of possibilities, as there are more search parameters. The search view looks like the view for adding new terms in Fig. 10 . If the purpose is to find all terms beginning with *blood*, for example, the



Term *hand*

Term: hand **Term source:** Oxford English Dictionary
Lexical fields:

- anatomy

Categories

- prim. med. meaning

Source	Entry	Year	Precis.	Ref
DOE	no entry			
MED	hond(e n.	1100	a.	OE
OED	hand n.1	825	c.	c825

Etymology
OE $hand, hond$.

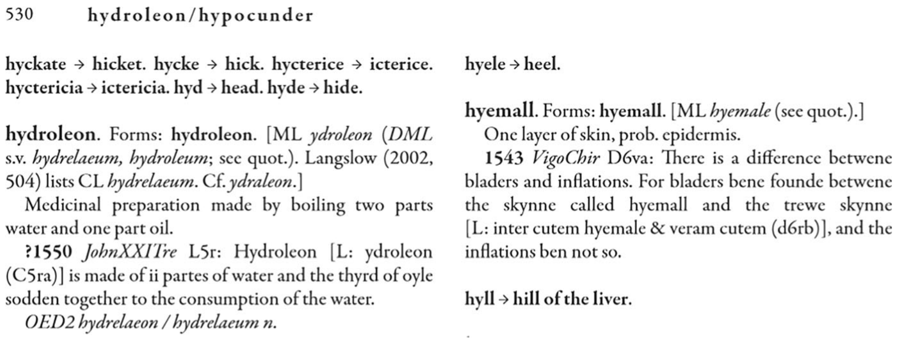
Senses

1. hand
2. arm and hand, arm below shoulder joint
3. (error) shin, leg between knee and ankle

Fig. 11 Searching for a term: result view

user types “blood*” in the “Term” box, the result being a list of 25 words and phrases including *blood*, *bloodiness of the eyes*, *blood leech*, *blood pissing*, and so on. A search for finding all the terms translated from foreign languages into English is carried out by typing “tr.” (i.e. translated) in the “Etymology” box. The relevant hits appear as a list, where clicking an item results in fuller information on it, found on a screen similar to the one used for browsing terms (Fig. 11). That screen is also used for adding a reference to books or articles which are relevant for understanding the medieval use of the term (the buttons at the bottom of the screen, not shown in Fig. 11, include “New reference text”).

One of the special features of the search for senses is that it does not yield just a list of the senses where a particular sequence of words occurs. By clicking any of those senses, the user gets a complete list of the terms expressing them. One learns, for example, that the sense ‘epilepsy’ is carried by no fewer than 29 terms, comprising *children evil*, *epilency*, *epilepsy*, *falling evil*, *falling sickness*, *foul evil*, *God’s wrath*, *holy passion*, *holy suffering*, and 20 other words and phrases.



530 hydroleon/hypocunder

hyckate → **hicket. hycke** → **hick. hycterice** → **icterice. hyctericia** → **ictericia. hyd** → **head. hyde** → **hide.**

hydroleon. Forms: **hydroleon**. [ML *ydroleon* (*DML* s.v. *hydrelaeum, hydroleum*; see quot.). Langslow (2002, 504) lists CL *hydrelaeum*. Cf. *ydraleon*.]

Medicinal preparation made by boiling two parts water and one part oil.

?1550 *JohnXXITre* L5r: Hydroleon [L: ydroleon (C5ra)] is made of ii partes of water and the thyrd of oyle sodden together to the consumption of the water.

OED2 hydrelaeon / hydrelaeum n.

hyele → **heel.**

hyemall. Forms: **hyemall**. [ML *hyemale* (see quot.).]

One layer of skin, prob. epidermis.

1543 *VigoChir* D6va: There is a difference betwene bladers and inflations. For bladers bene founde betwene the skynne called hyemall and the trewe skynne [L: inter cutem hyemale & veram cutem (d6rb)], and the inflations ben not so.

hyll → **hill of the liver.**

Fig. 12 Sample from a page of the printed dictionary

5.4 Generating the printed dictionary

The information stored in the database was selectively transferred to a printed version of the dictionary. Figure 12 presents the sections on the words *hydroleon* and *hyemall* in *DMV* (p. 530), showing how the entries look like. The headwords appear in bold type, as do their different spelling variants, here called "forms". The etymology of the word or phrase is given within square brackets. The meanings recorded from the corpus are listed in the paragraphs that follow. *Hydroleon* and *hyemall*, unlike many other names for medicines or body parts, are each used in only one sense in the texts analysed. In connection with individual senses, the reader finds quotations that illuminate the medieval use of the term. The source text and its date are specified. After the list of senses and quotations that pertain to them, any entries for the particular word in *DOE*, *MED*, and *OED* are indicated. One learns, for example, that *hydroleon* is given in the second edition of *OED*, but not in *DOE* or *MED*, whereas *hyemall* is recorded in none of the three dictionaries. As the last item in a *DMV* entry, there are references to articles and books in which the term has been discussed by other scholars. Sometimes, as for the two example words, none such have been found as yet.

DMV was published in two volumes, altogether circa 1300 pages. The database that provided the basis for the printed work contains more information than the latter, for which a selection had to be made of the 23,547 quotations and 37,596 variants (including 16,375 different variants) because of restrictions of space. The 2144 variants that the medieval writers signalled as belonging to a foreign language found little place in the published volumes. The total number of terms created in the process of compiling the dictionary is 12,605, coupled with 6168 senses. The resulting term-sense pairs ran to 14,985 items. Linked to individual terms, 1550 references to 177 different background sources are provided.

The printed dictionary called for some manual editing after the necessary items of information had been transferred into it from the database. Long quotations had to be shortened, keeping intact the central parts that threw light on the meaning and use of the term. In the list of combinations built round a particular headword, it was not infrequently the case that two or more compounds or phrases had an identical meaning. For example, the combinations listed under *blood* include 18 different ways of expressing the sense ‘blood become corrupt’ (e.g. *corrupt blood*, *evil blood*, *rotten blood*, *unlovable blood*, *unnatural blood*, *wicked blood*). In order to avoid unnecessary repetition in such cases, the sense was given in full only when it first appeared on the alphabetical list, with subsequent references to that definition.

6 *DMV* database and other dictionary formats

The digital application of *DMV* is implemented on top of a relational database. It has been argued, however, that relational databases are not suitable for storing and manipulating lexical data, but other approaches, such as markup languages, grammar-based models, and the feature-based model are better suited for lexical data (Neff et al. 1988; Véronis and Ide 1992; Ide et al. 1993). Markup languages maintain the original hierarchical structure of the lexicon entries, whereas in the context of the feature-based model, advanced manipulation needs such as recursive queries and flexible factoring are emphasized. In the following, we will introduce these models and investigate how our system takes such questions into account.

Among markup languages, SGML and its subset XML are probably the most widely used for representing lexical data. For example, *OED* and *MED* are based on them (nowadays XML). XML can also be seen as a successor of earlier grammatical approaches (Neff et al. 1988; Blake et al. 1992), because the structure of SGML/XML is usually described grammatically. Structurally, an XML document forms a hierarchy consisting of nested elements in terms of which the nested content of a lexical entry can be conveniently represented. XML does not require any predefined structure, which means that entries may vary from each other structurally. This has been an important property in developing digital dictionaries, because different sources (printed dictionaries) have different structures and ways of ordering the lexical entities. In other words, the entries can be first digitalized and the structures can be derived and integrated later on. In representing the structure, a grammatical approach (DTD, i.e. Document Type Definition) or the XML schema can be used. However, if the entries and the related elements vary from each other considerably in terms of structure, then the corresponding DTD is loose, i.e. it must allow structurally different kinds of documents. In other words, this kind of DTD does not determine the structure of the dictionary entries strictly. This has been observed for example in the developing of the electronic version of *OED* (Weiner 2009).

DTD and the XML schema can be seen as a schema-level description which corresponds to the relational schema in the sense that they all determine how the actual data are organized. An XML structure is a hierarchy, but lexical data also contains other kinds of relationships. For expressing such relationships, external references are used in XML and they can be seen to correspond to foreign key references in relational databases. In terms of external references, XML-based data can be reorganized into different kinds of hierarchies by XSLT, for example (Lemnitzer et al. 2013). The conceptual ER-model presented in this paper is more general than DTD and our database schema given in “Appendix”. Namely, the ER schema can also be transformed into DTDs (Suri and Sharma 2016). At the instance level, the data represented in relational databases can also be viewed hierarchically. The use case given in Sect. 5.4 demonstrates this, i.e. from the stored data a hierarchical view is formed for the printed dictionary. It would also be possible to construct other kinds of hierarchical views.

The feature-based model (Véronis and Ide 1992; Ide et al. 1993) or feature structure-based model (Trippel 2013), developed for advanced handling of lexical data, is based on logical attribute-value pairs that map entities to other entities. A value may be complex, which means that an attribute-value pair can represent a hierarchical structure. Attribute-value pairs can be manipulated by logical operations for uniting and restructuring them. This makes the feature-based model flexible in organizing data into different hierarchical views. The main motivation for developing the feature-based model has been the manipulation of (1) recursive nesting of entities, (2) factoring, and (3) different kinds of exceptions (Véronis and Ide 1992). We will next introduce how these questions relate to our design and implementation of *DMV*.

Recursive nesting of entities is also essential in *DMV*, because this kind of relationship appears in connection with senses, terms, and references. In the corresponding database schema in “Appendix”, recursive relationships are manipulated by recursive foreign key references. In this approach, a nested structure can be stored and then manipulated by a transitive closure algorithm. In *DMV* the transitive closure algorithm is implemented by the host language. Factoring, generally speaking, means a property for reorganizing lexical data into different hierarchies. As noticed in the context of XML above, the *DMV* data can be reorganized into different hierarchical views, although only one alternative is explicitly presented in the context of the printed version of *DMV*. The handling of exceptions has been seen as important for example in cases where the pronunciation of a word in a certain sense is different from the usual pronunciation (Ide et al. 1993). In our work, a similar situation occasionally arose in connection with questions of number, as illustrated in Sect. 4.2. In practice, such cases were handled by a separate value of an attribute in the Numbers table (see Appendix).

Although XML/SGML and the feature-based model possess their own indisputable advantages, a relational database is still a strong alternative for storing and manipulating lexical data. However, as also recognized by Vaquero et al. (2013), the use of the relational model requires a careful and thorough designing of the database structure. In the present paper, we have shown how this was done with the ER model. In our modelling of the lexical database for *DMV*, the standard historical dictionaries of English, in particular *MED* and *OED*, were very helpful in suggesting various concepts for inclusion. It proved necessary, however, to add a number of essential concepts of our own to the *DMV* database, without any equivalent in *MED* or *OED*. Such concepts relate in particular to the use of *DMV* for purposes of lexical research and include for example the word-formational/etymological category of the head word, information about the treatment of the word as a foreign word, and references to books and articles shedding light on the meaning and use of the term.

7 Evaluation and conclusion

We have now outlined the basic concepts of a database structure that was created for a specialized historical dictionary and described its most central functions from the user's point of view. As appears from the database schema in "Appendix", the survey has focused on the most essential features of the system, and many of the details had to be postponed for future discussion. The issues that have been addressed are likely to come up in the compilation of most dictionaries that aim at a wide coverage of the vocabulary of a special field during a certain period in history. The requirements posed on the database by a historical specialized dictionary turned out to be partly the same as for more extensive historical dictionaries, partly particular to works devoted to scientific or medical vocabulary.

In alphabetical dictionaries, the selection and arrangement of headwords is one of the most basic questions. In particular, the placing of extended units or phrases built around a specific word should follow principles that are clear to the users. In *OED* such units are given separately in a section titled "Compounds", which follows the meanings and uses of the simple word. Thus, for example, under the entry for *melancholy n.1*, the compounds listed are *melancholy-mad*, *melancholy madness*, *melancholy-purger*, *melancholy-sick*, and *melancholy water*. In the *OED* arrangement, the headword of the entire entry occurs as the first element of the compound. In the *DMV* database and printed work, a different approach was adopted. The compound or phrase was placed under its "parent term", grammatically the head of the whole unit modified by the other elements. *Melancholy madness* is thus found in the list of "Combinations" under the term *madness*. In this, *DMV* follows the policy of major modern medical dictionaries such as *Dorland's illustrated medical dictionary* (eds. Anderson et al. 2012) and *Stedman's medical dictionary* (eds. Stegman et al. 2006). In the former work, the two items forming a separate section under *melancholia* are *melancholia agitata/agitated melancholia* and *involutional melancholia*, the latter work giving *hyponchondiacal melancholia* and *involutional melancholia* in a corresponding position.

In the treatment of specialized terminology, the parent term approach has the advantage of giving the browser an immediate idea of how the phenomenon in question was classified in contemporary medical or scientific writings. The combinations placed under the term *blood* in the database comprise 89 phrases, and a mere glance at the list on the screen shows that one important basis of classification was the purity (e.g. *benign blood*, *natural blood*, *worthy blood*) or impurity (e.g. *bilious blood*, *corrupt blood*, *slimy blood*) of that bodily fluid. The browsing view also tells the user something about the frequency with which a term enters into multiword "extended terminological units" (Sager et al. 1980) and how complex those units are. The most prolific generator of combinations in our corpus was the term *water*, with no fewer than 611 extensions, mostly names of medicinal waters (e.g. *barley water*, *rose water*, *water of rosemary*). The most complex chain of terms and parent terms involves four levels: *vein*> *master vein*> *great master vein*> *great master vein of the liver*, the last three all signifying vena cava, a large vein opening into the right atrium of the heart.

As seen from the vena cava example and the 29 terms for epilepsy cited in Sect. 5.3, synonymy was rife in medieval medical works. There was no general agreement about the naming of individual organs, conditions, instruments, or medicines. Even today, it has been argued that complete synonymy is primarily found in technical language (Svensén 2009). Some writers observe that the phenomenon is in fact common in medicine (Pilegaard 1997; Landau 2001). For handling questions of synonymy, the database structure proved especially useful. The senses of the words and phrases were stored as independent units. Whenever a term appeared carrying a sense already recorded for some other term, it was simply linked to that existing sense. Because of this policy, the definitions of the medieval synonyms are worded identically in the database, which makes it easy to find all the different names that were applied to specific medical referents in the medieval material. Besides laying the foundation for a dictionary, the database can thus also be used as a thesaurus. It has to be stated, though, that the thesaurus aspect still calls for further developing. The senses were stored in the database as they came up in the corpus texts, and a total picture of the sense relationships was difficult to form at that stage. It was not infrequently difficult to decide what would be the parent sense, if any, for a new sense, and sometimes a suitable parent sense only appeared after the lower-level sense was already there. In a fully fledged thesaurus, individual senses are grouped under larger categories, and the principles of such categorization need further elaboration for Middle English and Early Modern English medical vocabulary. The monumental historical thesauri that exist for English (Kay et al. 2017; Roberts et al. 2017) will be helpful, but their information needs to be supplemented by the medical theory presented in the texts analysed for *DMV*. In the medieval classification of fevers, for example, the central criteria included the "humour" or fluid (blood, phlegm, choler, melancholy) thought to putrefy or increase excessively, causing body temperature to rise. Further distinctions were made according to the organs where the putrefaction or increase of the humour took place (see e.g. Bynum and Nutton 1981; Demaitre 2013). The role of the bodily humours or organs does not emerge from the treatment of the semantic category of fever in the entry 01.02.01.01.04.20 (n.) in Kay et al. (2017). An anonymous reviewer for *LRE* makes the pertinent suggestion that book illustrations and manuscript images might be considered for inclusion in the dictionary database in cases where they illuminate the medieval use of the term.

Sometimes the meaning of a medieval term remained uncertain or unclear in spite of a close scrutiny of the context(s) in which it appeared. The English physician John of Gaddesden refers several times to an ailment called *maras* in his *Rosa Anglica*, stating for example that "water of wormode is good...for dronkennes, for the sight and for the maras". It is possible that *maras* comes from Medieval Latin *marasmus* 'bodily wasting'. At the present state of knowledge, the sense of the word in *Rosa Anglica* cannot be determined with any degree of certainty and therefore carries a question mark in the database: '? wasting of body, emaciation'. By the completion of the project, many duplicated senses had arisen this way: one furnished with a question mark, the other without it. It would have been more economical to have just one sense and a separate marker for the uncertain instances. Similar considerations apply to errors in the corpus texts. For example, in one treatise the word *fistula* is erroneously used for the uvula. The relevant sense in the database is '(erroneous application) uvula'. For the sake of economy, again, it would be sensible to have just one sense 'uvula' and an additional identifier for the erroneous uses.

In historical dictionaries, the foreign and adapted forms of loanwords are often treated as separate entries, but not invariably so. *Virus* and *vire* 'pus from wound or ulcer' (< Latin *virus*) occur as two different headwords in *OED*, but *ranula* and *ranule* 'ranula, cyst under the tongue' (< Latin *ranula*) are both placed under the headword *ranula*. *Gangrena* and *gangrene*, as well as the earlier *cancrena* and *cancrene*, are all found under *gangrene n. (and adj.)*. In the *DMV* database and printed dictionary, the original and the modified versions are regularly listed as separate terms, with a cross-reference from one to the other. Unlike historical dictionaries, including the printed *DMV*, the database also contains all the words and phrases that the medieval writers assigned to a foreign language using formulae such as the ones cited in Sect. 4.1. Incompletely naturalized forms have been something of a headache for lexicographers (see e.g. Lewis 2007; Gilliver 2016), but their importance for the development of the lexical field should not be underestimated. Many of the terms that some writers singled out as belonging, for example, to Latin or French were used without any such comment by others. The information in the database enables a more comprehensive study of the history of the word or phrase especially in cases where the term has only gradually lost its foreign status in English medical writing.

References to background sources relevant to specific terms are a highly useful feature in historical specialized dictionaries such as *DMV*. Many of the lexical items occurring in medieval medical works go back to Latin or Greek, and scholars studying those two languages and medical treatises written in them have often discovered information that is missing from the English texts. Furthermore, medical historians sometimes disagree between themselves as to the exact referent of the medieval term, a good example being the *cells* or *cellules* of the brain (Norri 1998). For those consulting the dictionary, it is useful to be aware of the differing opinions. In the database, references to background sources were linked to terms, but in many instances they could also have been linked to senses because of the prevalence of synonymy in the corpus texts. If a certain article or book throws light on the meaning of a specific term, it is also likely to illuminate the meanings and uses of some of its synonyms. In future development of the database structure, the relationship between background references, terms, and senses deserves further attention.

The explicit classification of each headword into semantic, word-formational, and etymological categories is a feature that is usually absent from dictionaries. The presence of such data in *DMV* will be useful for lexicologists interested in the origins and use of medieval medical words and phrases, especially when entire lexical fields are being investigated. For example, it is possible to generate a list of all the terms that were created by English writers with the help of suffixes, or a list of all the English words that were used metaphorically to denote a sickness, body part, medicine, or instrument. Functions enabling a systematic pooling of certain types of lexical items proved most helpful in a recent article on translation from

Latin and French as a source of new medical terms in late medieval England (Norri 2017b). The first search criterion was the presence of the abbreviation "tr." (for "translated") in the "Etymology" field. Using the "Category" information in the database, the results so achieved were further categorized into simple terms, affix formations, and compounds and phrases, which were the three main groups discussed in turn in the article.

Any information system or software development project is finally evaluated through its use cases, i.e. how the particular application serves the requirements of users. In a software development project, the most serious problems typically follow mistakes in modelling an application domain or a universe of discourse. In other words, if there is an error in the underlying conceptual structure, this is reflected in the whole application and causes failures in various use cases. In the present electronic dictionary, all use cases succeeded well, which also means that the conceptual model and database works well. In practice, this is also due to the carefully designed iterative processing.

The database structure that was created for *DMV* could, with minor modifications, be applied to the historical study of vocabularies of other special fields than medicine.The history of scientific and medical writing is attracting increasing attention in many languages, and the questions that came up in the creation of our database are likely to emerge in other languages as well. A vast number of English medical terms, be they medieval or more modern, have their origins in Latin or Greek, and even Arabic is relevant for understanding the story behind some of them. Perhaps one day scholars interested in the development of medical vocabulary will have available to them a set of similarly structured databases for various languages, with mutual links between them. That would mean a significant widening of perspective not just in lexical matters but also in the history of various medical ideas.

The authors are currently looking into the possibility of producing a published version of the *DMV* database. At present, it is possible for interested scholars to gain access to the data for research purposes by writing to the corresponding author.

Acknowledgements The *Dictionary of medical vocabulary in English, 1375–1550* project was made possible by financial support from the Academy of Finland and the Finnish Cultural Foundation. We are greatly indebted to Paul Schaffner for a wealth of information on the digitization of the *Middle English dictionary*. Comments by Kimmo Rinteelä were most helpful at various stages of the writing process. We would also like to thank the three anonymous reviewers for their suggestions and observations.

Appendix: The database schema of *DMV*

See Fig. 13.

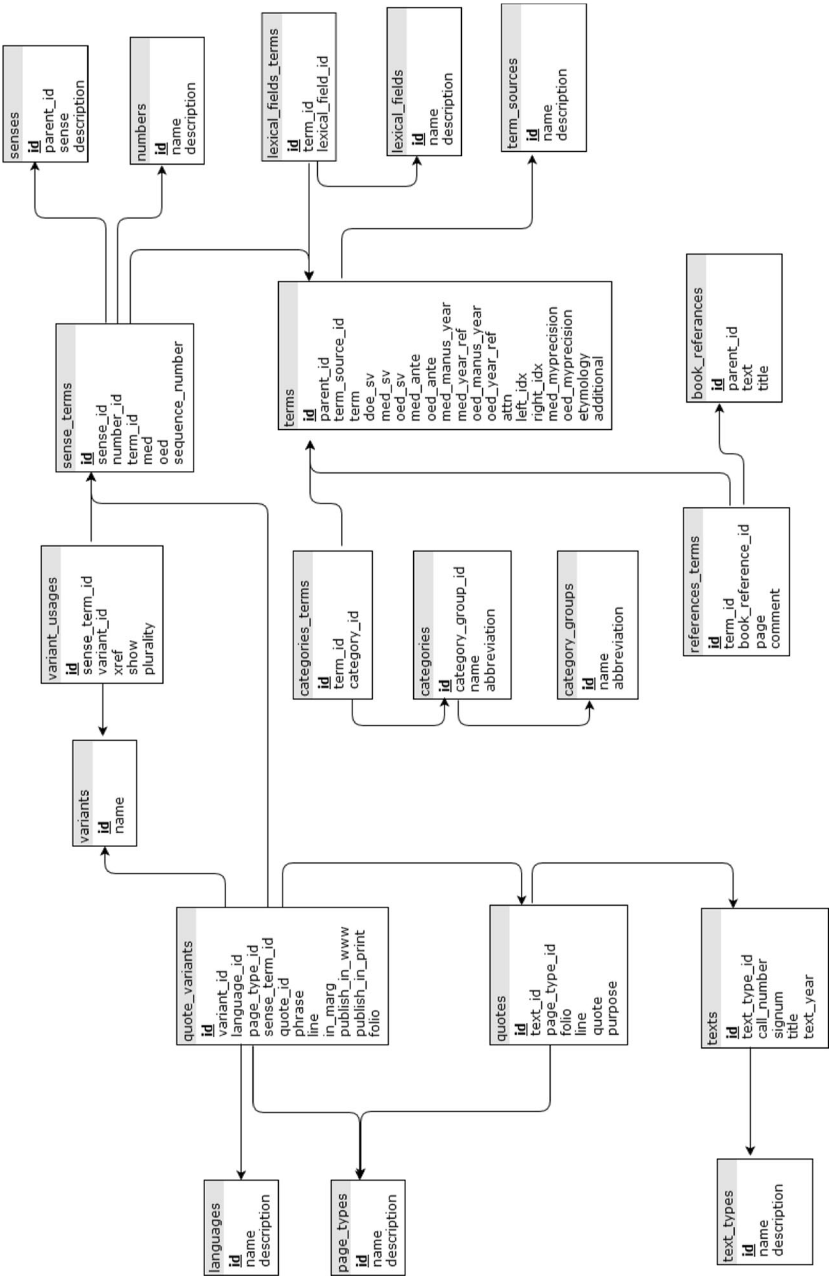


Fig. 13 Database schema

References

Anderson, D. M., Novak, P. D., Jefferson, K., & Elliott, M. A. (Eds.). (2012). *Dorland's illustrated medical dictionary* (32nd ed.). Philadelphia: Saunders.

Becker, H. (2016). Scientific and technical dictionaries; coverage of scientific and technical terms in general dictionaries. In P. Durkin (Ed.), *The Oxford handbook of lexicography* (pp. 393–407). Oxford: Oxford University Press.

Béjoint, H. (2010). *The lexicography of English: From origins to present*. Oxford: Oxford University Press.

Benskin, M. (1985). For wound in the head: A late mediæval view of the brain. *Neuphilologische Mitteilungen*, *86*, 199–215.

Blake, G. E., Bray, T., & Tompa, F. (1992). Shortening the OED: Experience with a grammar-defined database. *ACM Transactions on Information Systems*, *10*, 213–232.

Bynum, W. F., & Nutton, V. (1981). *Theories of fever from antiquity to the Enlightenment*. London: Wellcome.

CakePHP. (2018). CakePHP: An open source web framework. https://cakephp.org. Accessed 25 April 2019.

Cameron, A., Amos, A. C., Healey, A., et al. (Eds.). (1986). *The dictionary of Old English, A–I*. Toronto: Dictionary of Old English Project. http://tapor.library.utoronto.ca/doe/. Accessed 25 April 2019.

Cameron, M. L. (1988). On *þeor* and *þeoradl*. *Anglia*, *106*, 124–129.

Cawdrey, R. (1604). *A table alphabeticall, conteyning and teaching the true writing and vnderstanding of hard vsuall English wordes [etc.]*. London: Edmund Weaver.

Curzan, A. (2000). The compass of the vocabulary. In L. Mugglestone (Ed.), *Lexicography and the OED: Pioneers in the untrodden forest* (pp. 96–109). Oxford: Oxford University Press.

Demaitre, L. (2013). *Medieval medicine: The art of healing, from head to toe*. Santa Barbara: Praeger.

Dirckx, J. H. (1983). *The language of medicine: Its evolution, structure, and dynamics* (2nd ed.). New York: Praeger.

Elliott, L., & Williams, S. (2006). Pasadena: A new editing system for the *Oxford English dictionary*. In E. Corino, C. Marello, & C. Onesti (Eds.), *Proceedings of the XIIth Euralex international congress* (Vol. 1, pp. 257–264). Alessandria: Edizioni dell'Orso.

Foster, B. (1970). English 'jaw': A borrowing from French. *Neuphilologische Mitteilungen*, *71*, 99–101.

Gaskell, P. (1995). *A new introduction to bibliography*. Winchester: St. Paul's Bibliographies; New Castle, DE: Oak Knoll Press.

Getz, F. M. (1982). Gilbertus Anglicus Anglicized. *Medical History*, *26*, 436–442.

Gilliver, P. (2016). *The making of the Oxford English dictionary*. Oxford: Oxford University Press.

Gray, J. C. (1986). Creating the electronic new Oxford English dictionary. *Computers and the Humanities*, *20*, 45–49.

Hartmann, R. R. K., & James, G. (1998). *Dictionary of lexicography*. London: Routledge.

Haubrich, W. S. (2003). *Medical meanings: A glossary of word origins* (2nd ed.). Philadelphia: American College of Physicians.

Healey, A. (1985). The dictionary of Old English and the final design of its computer system. *Computers and the Humanities*, *19*, 245–249.

Hoare, M. R., & Salmon, V. (2000). The vocabulary of science in the *OED*. In L. Mugglestone (Ed.), *Lexicography and the OED: Pioneers in the untrodden forest* (pp. 156–171). Oxford: Oxford University Press.

Ide, N., Le Maitre, J., & Véronis, J. (1993). Outline of a model for lexical databases. *Information Processing & Management*, *29*, 159–186.

Jackson, H. (1988). *Words and their meaning*. London: Longman.

Kajaste, I. & Poranen, T. (Eds.). (2008). Software Projects 2008. University of Tampere, Department of Computer Sciences, Report D-2008-8.

Kay, C., Roberts, J., Samuels, M., Wotherspoon, I., & Alexander, M. (Eds.). (2017). *The historical thesaurus of English*. Version 4.21. Glasgow: University of Glasgow. http://historicalthesaurus.arts.gla.ac.uk/. Accessed 25 April 2019.

Keiser, G. R. (1978). Epwort: A ghost word in the Middle English dictionary. *English Language Notes*, *15*, 163–164.

Kurath, H., Kuhn, S. M., & Lewis, R. E. (Eds.). (1952–2001). *Middle English dictionary*. Ann Arbor: University of Michigan Press. http://quod.lib.umich.edu/m/med/.

Landau, S. I. (2001). *Dictionaries: The art and craft of lexicography* (2nd ed.). Cambridge: Cambridge University Press.

Lemnitzer, L., Romary, L., & Witt, A. (2013). Representing human and machine dictionaries in markup languages (SGML, XML). In R. H. Gouws, U. Heid, W. Schweickard, & H. E. Wiegand (Eds.), *Dictionaries: An international encyclopedia of lexicography. Supplementary volume: Recent developments with focus on electronic and computational lexicography* (pp. 1195–1209). Berlin and Boston: De Gruyter Mouton.

Lewis, R. E. (2007). *Middle English dictionary: Plan and bibliography* (2nd ed.). Ann Arbor: University of Michigan Press.

Logan, H. M. (1991). Electronic lexicography. *Computers and the Humanities*, *25*, 351–361.

McConchie, R. W. (1988). 'It hurteth memorie and hindreth learning': Attitudes to the use of the vernacular in sixteenth century English medical writings. *Studia Anglica Posnaniensia*, *21*, 53–67.

McConchie, R. W. (1997). *Lexicography and physicke: The record of sixteenth-century English medical terminology*. Oxford: Clarendon Press.

Neff, M. S., Byrd, R. J., & Rizk, O. A. (1988). Creating and querying lexical data bases. In *Proceedings of the second conference on applied natural language processing* (pp. 84–92). Stroudsburg, PA: Association for Computational Linguistics.

Norri, J. (1988a). A note on the entry *rede-wale* in the Middle English dictionary. *Notes and Queries*, *233*, 11–12.

Norri, J. (1988b). Notes on some corrupt passages in fifteenth-century medical manuscripts. *Neuphilologische Mitteilungen*, *89*, 320–323.

Norri, J. (1992). *Names of sicknesses in English, 1400–1550: An exploration of the lexical field*. Helsinki: Academia Scientiarum Fennica.

Norri, J. (1998). *Names of body parts in English, 1400–1550*. Helsinki: Academia Scientiarum Fennica.

Norri, J. (2004). Entrances and exits in English medical vocabulary, 1400–1550. In I. Taavitsainen & P. Pahta (Eds.), *Medical and scientific writing in late medieval English* (pp. 100–143). Cambridge: Cambridge University Press.

Norri, J. (2010). Dictionary of medical vocabulary in English, 1375–1550. In J. Considine (Ed.), *Current projects in historical lexicography* (pp. 61–82). Newcastle upon Tyne: Cambridge Scholars Publishing.

Norri, J. (2016). *Dictionary of medical vocabulary in English, 1375–1550: Body parts, sicknesses, instruments, and medicinal preparations*. Abingdon: Ashgate.

Norri, J. (2017a). The mystery of *mould* 'top of the head' in Middle English remedybooks. *Neuphilologische Mitteilungen*, *118*, 165–170.

Norri, J. (2017b). Translation from Latin and French as a source of new medical terms in late medieval England. *Romance Philology*, *71*, 563–622.

Osselton, N. E. (2009). The early development of the English monolingual dictionary (seventeenth and early eighteenth centuries). In A. P. Cowie (Ed.), *The Oxford history of English lexicography. Volume I: General-purpose dictionaries* (pp. 131–154). Oxford: Oxford University Press.

Pahta, P. (2011). Code-switching in Early Modern English medical writing. In I. Taavitsainen & P. Pahta (Eds.), *Medical writing in Early Modern English* (pp. 115–134). Cambridge: Cambridge University Press.

Pilegaard, M. (1997). Translation of medical research articles. In A. Trosborg (Ed.), *Text typology and translation* (pp. 159–184). Amsterdam: Benjamins.

Pollard, A. W. & Redgrave, G. R. (1986–1991). A short-title catalogue of books printed in England, Scotland, and Ireland and of English books printed abroad 1475–1640 (2nd ed. by W. A. Jackson, F. S. Ferguson, & K. Pantzer). 3 vols. Oxford: Oxford University Press.

Poranen, T. (Ed.). (2007). Software Projects 2007. University of Tampere, Department of Computer Sciences, Report D-2007-7.

Poranen, T. (Ed.). (2009). Software Projects 2008–2009. University of Tampere, Department of Computer Sciences, Report D-2009-6.

PostgreSQL. (2018). *PostgreSQL open source database system*. https://www.postgresql.com. Accessed 25 April 2019.

Roberts, F. (1980). *Medical terms: Their origin and construction* (6th ed. rev. by B. Lennox). London: William Heinemann Medical Books.

Roberts, J., Kay, C., & Grundy, L. (2017). *A thesaurus of Old English* (3rd ed.). Glasgow: University of Glasgow. http://oldenglishthesaurus.arts.gla.ac.uk/.

Sager, J. C., Dungworth, D., & McDonald, P. F. (1980). *English special languages: Principles and practice in science and technology*. Wiesbaden: Oscar Brandstetter.

Schäfer, J. (1989). *Early Modern English lexicography. Volume I: A survey of monolingual printed glossaries and dictionaries 1475–1640*. Oxford: Clarendon Press.

Simpson, J., Proffitt, M., et al. (Eds.). (2000). *The Oxford English dictionary* (3rd ed.). Oxford: Oxford University Press. http://www.oed.com.

Skinner, H. A. (1961). *The origin of medical terms* (2nd ed.). Baltimore: Williams & Wilkins.

Sonkusare, M., & Sahu, N. (2016). A survey on handwritten character recognition (HCR) techniques for English alphabets. *Advances in Vision Computing*, *3*, 1–12.

Stannard, J. (1982). Rezeptliteratur as Fachliteratur. In W. Eamon (Ed.), *Studies on medieval Fachliteratur* (pp. 59–73). Brussels: Omirel.

Stegman, J. K., Branger, E., Piper, T., et al. (Eds.). (2006). *Stedman's medical dictionary* (28th ed.). Baltimore: Lippincott Williams and Wilkins.

Suri, P., & Sharma, D. (2016). An algorithm for mapping ER schema in to XML DTD with recursion. *International Journal of Computer Applications*, *136*, 16–17.

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Thompson, P. A. (1992). The disease that we call cancer. In S. Campbell, B. Hall, & D. Klausner (Eds.), *Health, disease and healing in medieval culture* (pp. 1–11). New York: St. Martin's Press.

Trippel, T. (2013). Representing computational dictionaries in feature structure-based representation formalisms and typed feature logic. In R. H. Gouws, U. Heid, W. Schweickard, & H. E. Wiegand (Eds.), *Dictionaries: An international encyclopedia of lexicography. Supplementary volume: Recent developments with focus on electronic and computational lexicography* (pp. 1227–1234). Berlin and Boston: De Gruyter Mouton.

Vaquero, A., Alvarez, F., & Sáenz, F. (2013). Representing computational dictionaries in relational databases. In R. H. Gouws, U. Heid, W. Schweickard, & H. E. Wiegand (Eds.), *Dictionaries: An international encyclopedia of lexicography. Supplementary volume: Recent developments with focus on electronic and computational lexicography* (pp. 1209–1227). Berlin and Boston: De Gruyter Mouton.

Venezky, R. L. (1988). Unseen users, unknown systems: Computer design for a scholar's dictionary. *Computers and the Humanities*, *22*, 285–291.

Véronis, J. & Ide, N. (1992). A feature-based model for lexical databases. In *Proceedings of the 14th international conference on computational linguistics, COLING 1992* (pp. 588–594). Stroudsburg, PA: Association for Computational Linguistics.

Voigts, L. E. (1982). Editing Middle English medical texts: Needs and issues. In T. H. Levere (Ed.), *Editing texts in the history of science and medicine* (pp. 39–68). New York: Garland.

Voigts, L. E. (1984). Medical prose. In A. S. G. Edwards (Ed.), *Middle English prose: A critical guide to major authors and genres* (pp. 315–335). New Brunswick, NJ: Rutgers University Press.

Voigts, L. E. (1989). Scientific and medical books. In J. Griffiths & D. Pearsall (Eds.), *Book production and publishing in Britain, 1375–1475* (pp. 345–402). Cambridge: Cambridge University Press.

Voigts, L. E. (1995). Multitudes of Middle English medical manuscripts, or the Englishing of science and medicine. In M. R. Schleissner (Ed.), *Manuscript sources of medieval medicine: A book of essays* (pp. 183–196). New York and London: Garland.

Voigts, L. E., & Hudson, R. P. (1992). A drynke þat men callen dwale to make a man to slepe whyle men kerven him: A surgical anesthetic from late medieval England. In S. Campbell, B. Hall, & D. Klausner (Eds.), *Health, disease and healing in medieval culture* (pp. 34–56). New York: St. Martin's Press.

Voigts, L. E. & Kurtz, P. D. (2014). *Scientific and medical writings in Old and Middle English: An electronic reference*. Ann Arbor: University of Michigan Press.

Weiner, E. (2009). The electronic *OED*: The computerization of a historical dictionary. In A. P. Cowie (Ed.), *The Oxford history of English lexicography. Volume I: General-purpose dictionaries* (pp. 378–409). Oxford: Oxford University Press.

Woledge, B. (1970). English 'jaw': A borrowing from French. A footnote. *Neuphilologische Mitteilungen*, *71*, 467–469.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.