Unlocking the *OED*: the story of the *Historical Thesaurus of the OED*

Introduction

The *Historical Thesaurus of the Oxford English Dictionary* (*HTOED*) has a long and varied history, even if the story of its 45-year development is not quite as long as that of its principal 'parent', the *Oxford English Dictionary,* begun in 1879. This article, organized into four sections, outlines that history. Section 1 discusses the reasons for undertaking the project, the kinds of information it will yield, and some of the theoretical questions it may help to answer. Section 2 offers a narrative history of the project and its contributors from the initial collection of data to classification and the use of computers. Section 3 deals with the material itself, and especially with the issues involved in adapting data from the *Oxford English Dictionary* and dictionaries of Old English to the construction of a conceptually ordered historical thesaurus. Section 4 concludes with a description of the principles and procedures upon which the unique *HTOED* classification is based.

1. The purpose of *HTOED*

The main purpose of a modern thesaurus, of which the prime example is Roget's *Thesaurus of English Words and Phrases*, is to provide a collection of words 'arranged, not in alphabetical order as they are in a Dictionary, but according to the *ideas* which they express'.[1] It is as such, as a 'treasure-house' for those searching for words, for example speech-writers and crossword-solvers, that Roget's *Thesaurus* has become so well known through its many editions since his own time.

The present work is called a thesaurus because it shares the same feature of wordlists classified according to concept, idea, or (as in the case of concrete objects) nature or kind, and it can be used for the same purposes; but its main purpose is entirely different. It is the first *historical* thesaurus ever produced for any language. Put at its simplest, its purpose is to provide a detailed record of the English vocabulary from the earliest times to the present, with sufficient accompanying information that, for any given period in the past, the user should be able to ascertain the exact state of the vocabulary (i.e. the 'lexical system') which existed at that time. This degree of detail is possible only because its main source, the *Oxford English Dictionary* (*OED*), is the greatest single repository of facts about vocabulary available for any existing language. It presents the dates of currency for all the different meanings of each word, and thus includes, in addition to current meanings, all words that have become obsolete, and all obsolete meanings of words that still survive. This very full evidence of the *OED* covers the period from 1150 to the present, but its coverage for the Old English period (700–1150) is more selective. For that period, therefore, it has been necessary to add material from *A Thesaurus of Old English*.[2]

The purpose of this unique thesaurus, therefore, is to present in a more accessible form the vast amount of information which has hitherto been, as it were, locked away in the alphabetical order imposed by dictionaries—a problem pointed out as long ago as 1943 by the German scholar Walther von Wartburg.[3] Scholars who wished to reconstruct a lexical system of the past for English would accept the list of quasi- and near-synonyms given in Roget as representing the present, and trace back the history of each in the *OED.* They could perhaps find a few further relevant words in the *OED* quotations, but otherwise they would have no alternative but to go back to the actual texts of the period in question. If they relied on the evidence of Roget only, they would be omitting all the obsolete words included in the *OED*, as well as the relevant obsolete meanings of words that still survive. Thus, in the past, a number of monographs on individual words or groups of words have been published, but unless they are the result of very thorough and painstaking research in the texts of their period,[4] they are criticized because they omitted equally relevant words from that same conceptual system.[5] Electronic access to the *OED* has improved this situation somewhat, but until *HTOED* there has been no systematic conceptual presentation of the development of the vocabulary.

In comparison with the detailed and specialized work mentioned above, histories of the English vocabulary have tended to be more general, using criteria like etymology, types of word-formation, or loanwords as focal points for discussion. The obstacle to anything fuller has been the failure to recognize the primacy of meaning, and the analysis of meaning, as the

[1] From the preface to the 1852 edition, quoted in Susan M. Lloyd (ed.), *Roget's Thesaurus of English Words and Phrases*. London: Longman, 1982, xxi.

[2] Jane Roberts & Christian Kay with Lynne Grundy, *A Thesaurus of Old English*. King's College London Medieval Studies XI, 1995. Second edn, Amsterdam: Rodopi, 2000. Online version 2005.

[3] *Einführung in Problematik und Methodik der Sprachwissenschaft*. Tübingen, 1943. Revised edition with the collaboration of Stephen Ullmann, translated from the French edition by Joyce M. H. Reid, Oxford: Blackwell, 1969.

[4] E.g. A. Rudskoger, *Fair, Foul, Nice and Proper: A Contribution to the Study of Polysemy*. (Gothenburg Studies in English I.) Stockholm: Almqvist & Wiksell, 1952.

[5] E.g. A. Rynell, *The Rivalry of Scandinavian and Native Synonyms in Middle English*. (Lund Studies in English XIII.) Lund: Gleerup, 1948.

essential tool and criterion; for, in contrast to the closed systems of phonology and grammar, the huge open system of lexis demands nothing less than classification, by meaning, of the whole if one is to begin to understand the parts.

With this purpose in mind, the chronologically ordered lists of meanings presented in *HTOED* are intended:

(i) to give a history, with dates of currency, of the words used to express the concept or object stated in the list's heading, including losses, additions, and straightforward replacements; and

(ii) to act as a thesaurus for any period in the past, so that, for example, anyone wanting to know the range of words available to Shakespeare for a particular meaning can consult the appropriate timespan in the relevant section or sections.

It is the latter function that enables the user to return to the *OED* and gather fresh information about the character and history of all the words on each individual 'conceptual map', including changes of meaning, and redistributions of functions or meanings, especially those 'sideways shifts' of meaning which involve replacement by more than one word.[6] The proposed electronic linkage of *HTOED* and the *OED* will make it easier to gather such information in the future, since the user will be able to click on a word in *HTOED* and be taken straight to the *OED* entry, or vice versa.

In recent decades there has been considerable discussion about reasons for changes in the vocabulary. Some of the explanations advanced strike us as obvious, e.g. foreign invasion, the wealth of Latin loanwords at the time of the Renaissance, new inventions, or the simple replacement of one artefact by another. But other changes, and the reasons for them, are much more controversial. For example, opinions vary greatly on the relevance, as factors in vocabulary change, of features like homophony, near-homophony, phonetic and/or phonaesthetic suitability, and polysemy.[7] It is hoped that the presentation here of chronological lists (including obsolescences and therefore the choices of past speakers and writers) will provide today's scholars and critics with more of the evidence they need to answer such questions. It is also hoped that *HTOED* will offer fresh perspectives to scholars in other fields, for example anthropologists, historians, and others interested in cultural developments.

In addition to opening up the study of the successive and changing lexical systems of the past, the very fact of classification of meaning produces a whole new vertical perspective on each individual concept. Following the list of successive forms and their dates for the main concept, hierarchically organized sub-sections enable the user to compare the dates of closely related items. The results, especially where they include such tree-like structures, each with its own complete set of dates, can contribute substantially to determining the historical status of a concept with an observable past history. Such comparative procedures can be applied on the abstract side to the history of ideas, and on the more concrete to a large proportion of the remainder of the contents of *HTOED*, ranging from broad subjects like cultural and social history to others more specialized, such as military and domestic history.

2. The history of the project

Michael Samuels, then Professor of English Language at the University of Glasgow, announced the fact that his department intended to undertake production of a historical thesaurus of English at a lecture given to the Philological Society in 1965.[8] The original intention was that the work would be carried out by members of staff and postgraduate students, with individuals making contributions to both data collection and the development of the theoretical framework. Data collection came first, and each member of the team started to transcribe information from a volume of the first edition of the *OED*, using paper slips to record a word sense, its part of speech, its dates of recorded use, and any information contained in *OED* labels, such as 'figurative' or 'Philosophy', which might guide classification in future. Although the intention was never simply to produce a historical version of Roget's *Thesaurus*, its categories were used as a preliminary filing system; as the principle of classification primarily by semantic field developed, many Roget categories were virtually abandoned.

The members of the department who embarked on the work included Samuels, Leslie Blakely, Leslie W. Collier, John Farish, James Muir, and Jane Roberts; Roberts undertook to supplement the *OED*'s Old English materials. The first full-time postgraduate student was Irené Wotherspoon, who completed a thesis comparing a concrete area of lexis, the Body and its Parts, with the abstract field of Mental Pain.[9] The scale of the project was by then becoming apparent, and in 1969 a successful application for funding was made to the Leverhulme Trust, leading to the employment of Wotherspoon and Christian Kay as Research Assistants, mainly involved in collecting data. At the same time, substantial contributions to the archive of slips were made by volunteers, notably Frida Swanson, who made considerable inroads into the letter S, Frissy Peden, and Angus Somerville, who worked both at his home base at Brock University, Canada, and during research leaves spent in Glasgow. Data collection was also an integral part of the training of early postgraduate students, such as Elizabeth Donaldson, Ann Mackay Miller, who later became a Modern Humanities Research Association fellow on the project, working on Time and Change, and Freda Thornton, who was both a Research Assistant and a PhD student. Thomas Chase continued to work for the project after completing his Glasgow PhD and returning to the University of Regina; since he was working on Religion while Thornton tackled Good and Evil, there was considerable interchange of data and ideas between the two while they were doctoral students in Glasgow. Theses were also completed in London under Roberts' supervision, for example by Julie Coleman and Louise Sylvester. Titles of completed PhD theses or works based on them can be found in the Bibliography.

[6] For some results of such procedures, see Jeremy Smith, *An Historical Study of English: Function, Form, and Change*. London: Routledge, 1996, 135–139.

[7] See further: E. R. Williams, *The Conflict of Homonyms in English*. (Yale Studies in English, 100.) Yale University Press: New Haven, Conn, 1944; R. J. Menner, 'Multiple Meaning and Change of Meaning in English'. *Language* 21, 1945, 59–76; M. L. Samuels, *Linguistic Evolution*. Cambridge: Cambridge University Press, 1972, 67–77; Jeremy Smith, *op. cit.*, 120–123, 135–139; Dirk Geeraerts, *Diachronic Prototype Semantics*. Oxford: Oxford University Press, 1997; Philip Durkin, *The Oxford Guide to Etymology*. Oxford: Oxford University Press, 2009.

[8] See M. L. Samuels, 'The Role of Functional Selection in the History of English'. *Transactions of the Philological Society,* 1965, 15–40.

[9] I. A. W. Wotherspoon, 'A Notional Classification of Two Parts of English Lexis', M. Litt. Thesis, University of Glasgow, 1969.

Although now firmly established, in the coming years the project faced a series of intellectual, financial, and domestic challenges. The most dramatic of the latter was a fire in 1978, when the archive, by now amounting to many thousands of slips, survived only because it was housed in metal filing drawers within metal filing cabinets. Thereafter, the archive was microfilmed and new slips were completed in triplicate, with copies being stored at King's College London, where Roberts now had a lectureship, and in the Glasgow University Archive. When the Department of English Language moved to its current home in 1984, a former kitchen was converted into a fire-proof archive. Since great importance was attached to the completeness of the archive, both that move, and the two temporary moves which preceded it, placed a considerable strain on the team's organizational powers.

Fund-raising remained a perennial problem, producing many tense situations while people waited to hear whether a grant application had been successful and their jobs would continue. We became adept at dividing the project into manageable chunks which could be completed within the term of a grant. We practised economies in our use of resources: one colleague, perhaps with excessive zeal, worked out how many pages of the *OED* could be covered by a slip-maker with a single pencil (answer: 130). Renewal of the Leverhulme grant and a series of annual grants from the British Academy, plus funding from the University of Glasgow, enabled the Research Assistant posts to be maintained at various levels. When Kay became a lecturer in the department in 1979, her place as Research Assistant was taken by Freda Thornton. Other Research Assistants over the years included Lorna Gilmour, Ann Gow, Lesley Haughton, Cerwyss O'Hare, Liz Reay, and Judith Wood. Following Samuels' retirement from his chair in 1989, Kay took over as director of the project, assisted by Wotherspoon.

From 1981 to 1988 a new source of funding opened up, in the form of government-sponsored programmes for people learning new skills of editing and data entry in return for a stint of work on the project, starting with three trainees and peaking at nineteen. This development necessitated changes in the way the project was managed; a new stage of pre-classification was introduced, where trainees prepared sections of classification for future work by more experienced editors. It also saw the beginning of bulk input of data, and coincided with one of the major developments of the 1980s, the increased use of computers for storage and manipulation of the data.

The subject of computing had first come up in 1981 during talks with Oxford University Press about eventual publication of the project. It was made quite clear then that anything of such size and novelty could be printed only if it was handed over in electronic form, and the subsequent development of the *HTOED* database took place with that in mind. Following discussions with Glasgow University Computing Service, a database of 29 fields was designed and implemented by Alasdair Forsythe.[10] Already in the late 1970s, Roberts and colleagues at the King's College London Computing Centre had taken the first steps in computerizing the Old English materials, in order to provide a relatively small test corpus as a pilot study for the main project.

Forsythe's program proved robust throughout the entire period of data entry, but the rapidly moving pace of technology dictated further developments. Storage on the Glasgow University mainframe computer necessitated several changes of database, and the development of web technology opened up new possibilities for displaying and searching data. From the mid 1980s, work on developing versions of the database and data retrieval routines was carried out by Flora Edmonds, Joyce Farmer, Ann Miller, and Irené Wotherspoon.[11] Edmonds continued in the role of database officer until the end of the project, with help in the final stages from Jean Anderson and Marc Alexander. Administrative support was provided by Ian Hamilton. It has to be said that the role of the computer has been largely restricted to data entry, storage, and retrieval, although in the final stages, routines developed by Oxford University Press proved useful in identifying and correcting certain inconsistencies in the data. However, for the basic work of lexicography, i.e. extracting meanings and organizing them into categories, human input, working with paper slips, has been essential.

Basic slip-making from the first edition of the *OED* could have been completed by 1980, but before that a decision had been taken to include material from the Supplements to the *OED* published from 1972–86 and, later, from the second edition of 1989 and its Additions Series of 1993–97,[12] thereby enriching the project but also slowing it down. After that we decided to call a halt to data collection: it is hoped that revision of *HTOED* in line with the third edition of the *OED* will take place as that edition progresses.

Making a thesaurus is an endlessly circular process. From the late 1970s onwards, as described in section 4 below, we began to classify the data, starting with the more obviously discrete conceptual fields, such as Music and Food. Classification thus replaced slip-making as the team's principal activity; major grants to support both classification and the development of the database were received from the Leverhulme Trust, the Carnegie Trust for the Universities of Scotland, and the Arts and Humanities Research Board (now the Arts and Humanities Research Council). As the operation developed and the categories became less clear-cut, there was a good deal of movement of data between categories, with the result that sections completed in the 1980s built up a considerable backlog of material to be added. Nor was this necessarily the end of things, since the person revising a particular category might decide that some slips had been misdirected to it and send them on somewhere else. Slips from *OED2* also had to be added, often requiring matching with *OED1* senses. We could not therefore claim that the project was finished until the last slip was slotted into place in July 2008. Because of the complexity of the various stages in the editorial process, we have not, except in the case of published theses, attributed sections of classification to individual lexicographers: all sections, including thesis material, have been worked on by two, or usually more, people, in the twenty years since classification became the main focus of activity.

All in all, Samuels showed a certain prescience when he wrote in 1972:

[10] See C. J. Kay & T. J. P. Chase, 'Constructing a Thesaurus Database'. *Literary and Linguistic Computing*, 2, 3, 1987, 161–163.

[11] See Irené Wotherspoon, 'Historical Thesaurus Database Using Ingres'. *Literary and Linguistic Computing*, 7, 4, 1992, 218–225.

[12] *The Oxford English Dictionary*. 1884–1933, ed. by Sir James A. H. Murray, Henry Bradley, Sir William A. Craigie & Charles T. Onions; *Supplement*, 1972–1986, ed. by Robert W. Burchfield; second edn, 1989, ed. by John A. Simpson & Edmund S. C. Weiner; *Additions Series*, 1993–1997, ed. by John A. Simpson, Edmund S. C. Weiner, & Michael Proffitt; third edn (in progress) *OED Online*, March 2000– , ed. by John A. Simpson. Oxford: Oxford University Press.

> The production of such a thesaurus is arduous. Every attested meaning, past and present, must be semantically analysed and classified, and this can be achieved only by conventional methods, not by computer. It is at present being attempted for English only, and, from experience so far gained, will be a lengthy task.[13]

3. The data

As noted above, the primary source of data for the project was the second edition of the *OED* and its additions, incorporating the first edition and its supplements. For the Old English period, these data were augmented by slips covering the recorded vocabulary, whether obsolete by 1150 or not. These Old English slips were compiled by Roberts, using Clark Hall's *A Concise Anglo-Saxon Dictionary*[14] as an initial word list, and Bosworth Toller's *An Anglo-Saxon Dictionary*[15] for its supplementation and for fuller exemplification of senses. Account was also taken of materials from the new Toronto *Dictionary of Old English* as they became available.[16] As Roberts and Kay worked on the classification of these materials, it became clear that they would make a useful resource for scholars in their own right, hence the publication in 1995 of *A Thesaurus of Old English* (*TOE*)[17]—a rare case of the child preceding the parent. Data from *TOE* are included in *HTOED*, but without the length marks normally added in Old English texts or the flags used in *TOE* to indicate distribution and frequency. Words from *TOE* appear either as freestanding Old English forms, labelled OE, or are linked to their modern descendants as listed in the *OED*. Thus *mother<modor OE–* represents the modern English word *mother*, derived from OE *modor*, and in continuous use thereafter. Since one of our main interests was to see how many Old English words mapped on to the *OED* entries, we may have been over-generous in our distribution of Old English forms. On the other hand, an interesting by-product of this operation has been the discovery of possible links spanning centuries of unrecorded use, as when an apparently obsolete Old English word resurfaces in a nineteenth-century dialect dictionary.

Although we have relied heavily on the *OED*, we have not followed it slavishly. In a paper outlining the relationship between the *OED* and *HTOED*, Kay and Wotherspoon wrote:

> Given the different purposes and design of dictionaries and thesauri, it is not surprising that difficulties should arise in transforming one into the other. The basic aim of the *OED* is to give maximum information about the development and use of individual word forms. HTE [= Historical Thesaurus of English, the project's working title], on the other hand, aims to group together words which share one or more features of their meaning, thus using a broader brush to maximize semantic information. The thesaurus slip-maker is not mechanically transferring data from one format to another, but is continually making judgements about the appropriateness of that information for his or her purpose. Thus, the original *OED* editor may have felt that information about a certain word suggested a division into a certain number of senses and proceeded accordingly. The thesaurus editor, on the other hand, may feel that these divisions are either too broad, and thus miss nuances of meaning that s/he might wish to see represented, or, more usually in our case, too narrow, resulting in several appearances of the same form within a single semantic category. The fact that such problems occur does not mean that either editor is 'wrong' in his or her decisions, but is an inevitable outcome of data intended for one purpose being adapted for another.[18]

Thus, although the *OED* sense divisions are generally followed, there is not total isomorphism between the data in *HTOED* and *OED*. Nor do we claim to include the entire contents of the *OED*. Where a word generates a large number of phrases and compounds, we have usually omitted the most obviously transparent. We have also been selective with derived forms, such as the almost limitless formations with prefixes such as *dis-* or *un-*.[19] There are sometimes apparent absences of words from *HTOED* lists, not because the *OED* does not include them, but because (especially in the case of verbal nouns and participial adjectives) it does not include a quotation in the sense needed for that particular list. This applies especially where the *OED* has a blanket entry to cover 'action of the verb in its various senses' without specifying or distinguishing the senses, for example 'sagging' verbal noun, where the single entry contains citations which could potentially be linked to any of the fourteen meanings of the verb. The result is that it is quite often difficult to be sure which sense each form is to be assigned to, and it seems safer to omit the form altogether; conceivably the *OED* editors themselves were not sure. A source of unevenness for the Old English data, again relating to participles and participial adjectives, is that these are represented only very sparsely among the headwords in Clark Hall and Bosworth Toller. Compounds like *landagende, swiðhycgende* are included; simplices like *reocende* appear only sporadically. The result is that the proportion of these forms in the adjective lists is much lower for Old English than for Middle and Modern English.

Similar situations can arise with adjectives and adverbs. The senses of an adjective may be less finely differentiated than those of the noun from which it derives, or the adverb senses from those of the parent adjective. The slip-maker then has to decide whether to put the adjective in a more general category than the noun, or to select the citations appropriate to each noun as far as this is possible, and bearing in mind that the *OED*

[13] *Linguistic Evolution*, 180.

[14] J. R. Clark Hall, with supplement by H. D. Merritt, *A Concise Anglo-Saxon Dictionary.* Fourth edn. Cambridge: Cambridge University Press, 1960.

[15] Joseph Bosworth & T. Northcote Toller (1882–98) with supplements by Toller (1908–21) and A. Campbell (1972), *An Anglo-Saxon Dictionary.* London & Oxford: Oxford University Press.

[16] A. F. Cameron, A. C. Amos, A. diP. Healey, Sharon Butler, Joan Holland, David McDougall, & Ian McDougall (eds), *Dictionary of Old English (DOE).* (Published for the Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.) Toronto: PIMS, 1986– .

[17] Roberts & Kay, *op. cit.*

[18] 'Turning the Dictionary Inside Out: some issues in the compilation of a historical thesaurus'. *A Changing World of Words: Studies in English Historical Semantics and Lexis*, Javier E. Diaz Vera (ed.). Amsterdam: Rodopi, 2002, 109–135.

[19] For an interesting account of the *OED*'s issues with the latter, see Peter Gilliver, 'The Great Un-crisis: an unknown episode in the history of the *OED*'. *Words and Dictionaries from the British Isles in Historical Perspective*, John Considine & Giovanni Iamartino (eds). Newcastle: Cambridge Scholars Publishing, 2007, 166–177.

editors had access to discarded examples as well as those which ended up in print. On the whole we have gone for the latter solution rather than simply repeating the adjective entry in all the possible categories.

Derived verbal forms also raise issues of transitivity. In the *OED*, transitivity is rarely (if ever) specified for verbal nouns and participial adjectives, but for *HTOED* purposes we needed to know which form of the verb they were attached to if we were to take their dates into consideration. There are also issues, not just for dictionary or thesaurus makers but for grammarians generally, in the wide grey area between an intransitive verb standing alone and a transitive verb with a direct object—a cline of transitivity including such following constructions as indirect objects, infinitive phrases, clauses, and prepositions. On the whole we have tended to follow more recent *OED* practice in labelling a verb with any kind of object as transitive. The problem, however, is compounded where there are Old English verbs, since objects may occur in different grammatical cases, or there may simply not be enough evidence to support a decision. In that case, verbs are entered without specifying transitivity, giving us three basic categories: vi. (intransitive), vt. (transitive) and v. (verb). There are also three minor categories: v. impers. (impersonal), v. pass. (passive), and v. refl. (reflexive).

As with any large project conducted by many hands over a long period of time, there have been some variations in practice in both slip-making and categorization. Originally, for example, we intended to exclude two classes of words where the coverage of *OED1* was known to be patchy: later dialectal words and words from recent technological and scientific fields. Both of these initial decisions were overtaken by events, principally the much greater coverage of scientific registers in *OED2*, reflecting a more widespread interest in things scientific.[20] Many dialect words crept in simply because the compilers could not resist them, but also because of increased interest in linguistic variation. The basic meanings of grammatical words, such as prepositions and conjunctions, are usually represented, but not at the level of detail given in the *OED*.

In matters of spelling, we have followed *OED2* headwords, including any variant spellings that occur there, and occasionally including further variants if they are significantly represented in particular senses. Capitalization is less straightforward. The preface to *OED2* tells us that:

> In the first edition of the *OED*, every main headword was given a capital initial, regardless of whether the word was normally so written. Most derivatives, and many combinations, were also capitalized. The Supplement, in accord with modern lexicographical practice, abandoned this convention, giving a capital only where that is the normal spelling. This edition follows the Supplement's practice. For many words capitalization varies, either at different dates or in different senses. Because its convention disguised the problem, the *OED* often did not indicate the prevailing or preferred style. Where the intentions of the first edition were not deducible, as often with rare and obsolete words, decisions about capitalization were made on the basis of the printed quotations or analogy with similar and related words, or both.[21]

In general, we have followed this practice, treating capitalization of each sense of a word according to the information available in the *OED* headwords and citations. Sometimes this conflicts with the *HTOED* style of using upper case initials for main category headings and lower case for subordinate ones. For proper names, we have allowed capitalization to override this style, but in one particular case, the use of Latin taxonomic terms in subordinate scientific categories, we have retained lower case initials.

Where appropriate, we have included *OED* labels indicating features such as style (the abbreviated forms used in *HTOED* are given in square brackets), for example 'slang' or 'ironic' [*iron.*]', and provenance (e.g 'dialectal' [*dial.*] or 'South African' [*S. Afr.*]). However, we have omitted such labels where the fact of classification makes them redundant; we do not, for example, label a word 'Physics' in the category of that name. Later Scots words are usually labelled as such, but for older words, where there is an initial mixture of Middle English and Older Scots citations, labelling was often deemed unnecessary. One of the most problematic labels has proved to be 'figurative' [*fig.*]. The huge amount of research into metaphor in recent years has focused attention on the extent to which the abstract lexis is derived from the concrete, i.e. is inherently metaphorical.[22] In a category such as Pride, when one's perception of metaphor is sharpened, one begins to wonder why *upstage* is marked as figurative, while *condescend* is not. The criterion presumably has to be whether the metaphor is sufficiently fresh to be perceived as such by users of the language.

HTOED has various ways of indicating currency. Words with an *OED* final date after 1870 are usually considered to be actually or potentially still in use, and the closing date is replaced with a dash, as in *aunt 1297–*. In cases of uncertainty, we consulted available sections of *OED3*, online corpora, and two desk dictionaries, *The Concise Oxford Dictionary*[23] and *The Chambers Dictionary.*[24] Where uncertainty persisted, we retained the last recorded date. For cases where there were doubts about continuous currency, usually defined as a gap in citations of around 150 years without any obvious reason, we have replaced the customary dash between dates with a semi-colon. Thus, *great-aunt 1656;1870–* indicates that the word was first recorded in 1656, but not found again until 1870, whereas *nephew 1494–1585* (in the sense of 'niece') indicates that the word appears to have been current between those dates.[25] Where a word (or sometimes a date) is recorded only in a dictionary

[20] See Michael Rand Hoare & Vivian Salmon, 'The Vocabulary of Science in the *OED*'. *Lexicography and the OED: Pioneers in the Untrodden Forest*, Lynda Mugglestone (ed.). Oxford: Oxford University Press, 2000, 156–171.

[21] From the *OED2* preface online, point D5. http://oed.com/archive/oed2-preface/intro-features.html

[22] The classic text, which inspired much subsequent work, is George Lakoff & Mark Johnson, *Metaphors We Live By*. Chicago: University of Chicago Press, 1980.

[23] R. E. Allen (ed.), *The Concise Oxford Dictionary*. Eighth edn. Oxford: Clarendon, 1990.

[24] Robert Allen & Catherine Schwarz (eds), *The Chambers Dictionary.* Edinburgh: Chambers, 1998.

[25] Durkin, *op.cit.,* chapters 3 and 8, has an interesting discussion of how, if at all, we can determine whether words and meanings with widely separated citations are cases of re-invention at different periods ('polygenesis') or simply victims of a defective record. Some light may be shed on such matters when words are examined in semantic categories in *HTOED*.

or similar work, we have used the label 'Dictionary' [*Dict.*], to indicate doubts about its general currency. In all cases, a certain amount of lexicographical discretion was allowed.

4. The classification

As data collection proceeded, the attention of the team turned increasingly to the immense task of devising a system of classification which would do justice to such a large and varied body of material. Such a system had to be flexible enough to accommodate changes in the vocabulary over the years and the cultural changes they reflected. We were also keen to include a much greater degree of semantic discrimination than is found in most thesauri. While the resultant framework inevitably coincides with those of other thesauri at certain points, as a whole it offers a uniquely detailed system for semantic classification.

At the root of the *HTOED* system of classification is the contention that, within certain limits, each section should be allowed to develop its own semantic profile. However, as the body of data grew, it was felt that a high-level overall structure should be developed into which the categories assigned to individual classifiers could eventually be slotted. Samuels and Kay undertook the task of identifying key components in the *OED* definitions which would form the basis of major sections.[26] Further work led to a set of 26 major categories, forming a framework for *HTOED* as a whole. Within this framework there is provision for seven main category levels and five subcategories, in a taxonomy which begins with the most general ways of expressing a concept and moves hierarchically downwards to the most specific. In linguistic terms, we are applying an organizing principle of hyponymy, encapsulating relationships such as 'type of' or 'part of'. In the present volume, numbered category headings begin new paragraphs of text, while subcategory headings appear in italic within the paragraph. All twelve levels are sometimes used in classifying the objects of the material world, where a good deal of detail can be specified, but are rarely needed in the broader divisions of abstract categories.[27]

There was much discussion of how parts of speech should be represented within this structure, with arguments put forward both for allowing semantics to override grammar, resulting in intermingled parts of speech, and for identifying a leading part of speech in each category, so that concrete categories would normally begin with nouns but more abstract ones, such as Thought or Goodness, might begin with verbs or adjectives respectively. In the end, in the interests of ease of use, it was decided to maintain a consistent pattern of Noun, Adjective, Adverb, Verb within each category and subcategory, followed by Phrase, Interjection, Conjunction, and Preposition.

As it now stands, *HTOED* is presented in three major sections, reflecting the following broad semantic divisions:

I The external world
II The mental world
III The social world

This tripartite division reflects the fact that, for English in the British Isles, we are dealing with a world-view, or set of world-views, recorded over a period of about 1300 years, but often incorporating much earlier views. Our historical perspective enabled us to tackle one of the key problems of any system of classification: where do you start? For Roget, the answer was to start with abstractions, notions such as relationships of similarity, comparability, etc., which would inform later sections. For *HTOED*, the solution was the opposite: to start in Section I with the most readily observable phenomena of the universe, the land, sea, sky, etc., followed by living beings, their characteristics and physical needs. At the end of this section, which has by far the largest number of meanings, come attempts to quantify and interpret the world through systems such as space, time, and measurement. Section II, the smallest of the three, presents the vocabulary of mental processes, such as Perception, Emotion, Will, and (perhaps more marginally) Possession. Here there is a logical progression, since much of the lexis of this section derives metaphorically from that of Section I.[28] Now that the entire thesaurus is available, and can be examined as a whole in this volume, more work can be done on sections where substantial transfer of vocabulary between categories suggests possible metaphorical or other links.

Section III deals with the vocabulary of people as they organize into social groups, develop systems such as law and morality, exploit the environment, communicate, and enjoy themselves. It contains the largest number of categories (as opposed to meanings), reflecting huge changes in social organization and activities over the years; categories such as Leisure and Entertainment, which are tiny in *TOE*, have grown almost beyond recognition in the modern period, reflecting changing lifestyles. The same, of course, applies to major scientific categories in Section I, such as Physics or Medicine.

Categories such as Law, Medicine, or Will, described loosely as semantic fields, formed the basic unit of classification, being of a size (between about 10,000 and 20,000 slips) which an individual could reasonably be expected to tackle, and which could be expected to yield interesting results. The decision to base the classification on such fields led to major departures from Roget's framework, particularly in the transfer of categories from Section I to Section III. Sometimes, the individual produced a detailed classification of a particular area, which could later be added to, as in the case of work done by postgraduate students[29] and research assistants, and, as the project became better known, by colleagues from other universities, notably Reinhard Gleissner of Regensburg University, Hans Peters, now at the University of Dortmund, and Angus Somerville of Brock University, Canada. In other cases, trainee lexicographers did preliminary sorting, which was later refined by an experienced editor.

It was acknowledged from the start that each section should be allowed to develop its own structure. Within the general taxonomic framework, classifiers were given a free hand, being told simply to 'sort, sort, and sort again' until an acceptable

[26] C. Kay & M. L. Samuels, 'Componential Analysis in Semantics: Its Validity and Applications'. *Transactions of the Philological Society*, 1975, 49–81.

[27] See Christian Kay & Irené Wotherspoon, 'Semantic Relationships in the Historical Thesaurus of English'. *Lexicographica* 21, 2005, 47–57.

[28] For an extended examination of *HTOED* data in such terms, see the thesis by Reay cited in the Bibliography and Kathryn Allan, *Metaphor and Metonymy: A Diachronic Approach*. Oxford: Blackwell, 2008.

[29] See, for example, the theses listed under Chase, Coleman, Sylvester, and Thornton in the Bibliography, each of which contains a classification as well as analysis of various linguistic aspects of the data. The thesis by Chase also made a significant contribution to the development of *HTOED*'s notation.

structure emerged; in other words, the classification was 'bottom up' from the data rather than imposed 'top down'. Our aim was to produce a folk taxonomy, informed by what Hallig and von Wartburg describe as 'naïve realism', setting forth 'the intelligent average individual's view of the world, based on pre-scientific general concepts made available by language'.[30] However, since in some sections, such as Animals and Plants,[31] we found that an established scientific taxonomy was the best way of dealing with some of the data, we ended up with what we describe as a 'modified folk taxonomy', where the naïve view may be combined with one that is more informed. From this point of view, the ideal classifier is a person who combines linguistic sophistication with a degree of appropriate subject knowledge, and many classifications were assigned with this in mind. Using scientific terminology also helped to solve the problem of devising explanatory headings for complex scientific categories within the 50-character limit imposed by the database.

There is obviously no single way of structuring a thesaurus, as comparative studies of the semantic systems of the relatively few existing thesauri have shown.[32] Nor is any semantic category likely to be wholly clear-cut. Initially, it might seem that a field like Music or Religion, two of the earliest to be tackled, has well-defined content, but even here problems arise, such as where to assign Religious Music. A more serious problem for *HTOED* was how, or indeed whether, to draw a boundary between Religion and cognate areas which might now be excluded from it, such as Magic and Witchcraft. In the end a decision was made to set up two fields, one for all supernatural beings and manifestations, regardless of originating creed, and the other for organized religion; experimentation while compiling *TOE* confirmed that such a model was appropriate for the long historical sweep of *HTOED*. These two categories are quite widely separated in the thesaurus as a whole, appearing in Section I (as an attempt to explain the world) and Section III (as a social construct) respectively.

The need to make such decisions recurred throughout the work. In some cases, the sheer weight of vocabulary simply overwhelmed the taxonomy, with a category that is properly a subcategory rising in status because its degree of lexicalization reflected its considerable degree of importance to speakers of the language. Thus, for example, historic and important sports, such as cricket, football, or baseball, have their own categories and arrays of subcategories. In all cases, it must be stressed that we are classifying the language used to discuss a topic, not the subject matter itself, which may sometimes lead those who are knowledgeable about a topic to find a particular category defective. Our category of Countries, for example, reflects the priorities of a dictionary and would not satisfy a geographer; it indicates the various ways in which people have referred to parts of the world but is not an encyclopedic list of polities.

[30] Cited in Ullmann, *Semantics*. Oxford: Blackwell, 1962, 255.

[31] In the case of Plants, both types of taxonomy were attempted. See the study reported in Cerwyss O'Hare, 'Folk Classification in the HTE Plants Category'. *Categorization in the History of English*, Christian J. Kay & Jeremy J. Smith (eds). Amsterdam: John Benjamins, 2004, 179–191.

[32] See, for example, Andreas Fischer, 'The notional structure of thesauruses' in Kay & Smith (eds), *op. cit.*, 2004, 41–58; Werner Hüllen, *A History of Roget's Thesaurus: Origins, Development, and Design.* Oxford: Oxford University Press, 2004.

Within each heading in *HTOED*, meanings are grouped according to a loose principle of synonymy. There is no claim that these words are exactly synonymous, i.e. could replace one another in all contexts (if such a condition exists), but rather that they share enough of their meaning to be classified together. Although the project was started before the current cognitive semantics paradigm became dominant, that paradigm has retrospectively proved sympathetic to the problems involved in categorizing large quantities of lexical data. The development of prototype theory, which allows for fuzzy sets containing both good and less good examples of the central concept, challenges the either/or basis of Aristotelian category assignment and liberates semanticists from a narrow notion of synonymy as an organizing principle.[33] *HTOED*'s synonym groupings are prototypical in nature, with a clear core of obvious members shading off into the less obvious, and ultimately into subordinate or cognate categories.

Such flexibility is especially essential for historical semantics, allowing connections to be made and new systems identified. Thus, for instance, where a form occurs in category (a), and also appears, usually with a later starting date, in category (b), there may be evidence of gradual semantic change through a polysemous chain of meaning. One example among many is the adjective *sensitive,* which first occurs in the fifteenth century in 01.03.00|06 Physical sensibility in the sense 'Having the function of sensation'. Several meanings related to physical sensation develop within that section, until in 1816 the word is recorded with reference to mental sensation, classified in 02.02.08.01 Emotion. If we examine the latter category and those surrounding it, we find other words with a similar transfer of meaning from concrete to abstract, such as *soft, tender*, *sensible*, and various forms of the verbs *touch* and *impress.* Interestingly, the expression *sensitive plant* has made a similar journey, this time from a botanical term for a literally irritable plant to a way of describing a sensitive person. However, the most recent meaning of *sensitive*, as in 'sensitive information', has made a lonely journey to 02.01.12.10.01.01 Knowledge, subcategory 02 *not printed/published*, where no other words of comparable origin are to be found.

The other common organizing principle of thesauri, antonymy or oppositeness of meaning, has proved less suitable for systematic use in *HTOED*, since oppositions vary in both nature and extent. Where there are substantial categories containing obvious oppositions, as in Love/Hate or Pain/Pleasure, they are generally placed together, but in other categories, such as Truth, there can be a progression of meaning covering several oppositions, in this case moving from Validity through Truth, Sincerity, Falsehood, and Error to Deceit. Where opposites are too few to merit a separate category, they are usually classified after the meaning they negate. The same principle applies to phrases. If there are only a few, they are classified with the appropriate part of speech; where there are many, they are grouped in a separate category of phrases.

Each category and subcategory in *HTOED* has a modern English heading; wherever possible, the keyword in the heading is drawn from the category, but sometimes a word of more general scope is used. Sometimes there will be a list of subcategories without any words of more general meaning to form a superordinate category. In such cases, an empty heading may be supplied, as in the example below from the section on butterflies in the Animals category:

[33] See, for example, John Taylor, *Linguistic Categorization.* Third edn. Oxford: Oxford University Press, 2003.

01.02.06.13.11.08.10 | 04.03.07 *genus Argynnis*
01.02.06.13.11.08.10 | 04.03.07.01 *member of/fritillary* turkey-egg, fritillary

Here, there are no actual exponents of the first subcategory, which is included in order to maintain the taxonomic structure. Below it is a further subcategory, reading 'member of the *genus Argynnis*', where we find two expressions for the fritillary butterfly.

Recurrent headings may be given in abbreviated form, as in '*not* (having the quality described above)', '*one who* (performs a certain action)' or '*instance of* (a particular condition)'.

Throughout *HTOED*, headings are intended to link in this fashion, so that reading back from the lowest to the highest level will construct an approximate definition revealing the layers of taxonomic structure with which a meaning engages. Indeed, it would be possible to produce a novel kind of dictionary by this process, turning inside out the thesaurus which was constructed by turning the dictionary inside out, but that would be a project for another day.

Christian Kay Jane Roberts Michael Samuels Irené Wotherspoon