n

Kay, C., and Alexander, M. (2015) Diachronic and synchronic thesauruses. In: Durkin, P. (ed.) The Oxford Handbook of Lexicography. Oxford University Press. ISBN 9780199691630

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

http://eprints.gla.ac.uk/70403/

Deposited on: 19 January 2016

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

24. Diachronic and Synchronic Thesauruses

Christian Kay and Marc Alexander, University of Glasgow

24.1. The Thesaurus

A close sibling of the dictionary, thesauruses are those works of lexicographical reference which present lexical facts with semantic domains as their core organizational principle, rather than in alphabetical arrangement. Far from being a simple repositioning of existing dictionary entries in topical order, thesauruses are the oldest recorded form of lexicographical work, and blend facts about the language with facts about the world in which the language is used. As a result, the classification system at the heart of a thesaurus represents a synthesis of the conceptual vocabulary of a language, and of the relative place of each concept with regards to the wider conceptual system of which it is a part.

24.1.1. At the End of the Alphabet

While it may be natural for many present-day readers to consider a thesaurus as an accessory to a dictionary, given the dominance of the latter format's alphabetical order, this situation is mostly due to the dictionary system's ubiquity in recent years. By contrast, such was the historical dominance of the meaning-based ordering of information, the dictionary's alphabetical arrangement was so strikingly novel in its early publication history that Robert Cawdrey in his 1604 *A Table Alphabeticall* felt it necessary in his prefatory note "To the Reader" to explain this system, in a notoriously awkward passage:

> If thou be desirous (gentle Reader) rightly and readily to vnderstand, and to profit by this Table, and such like, then thou must learne the Alphabet, to wit, the order of the Letters as they stand, perfecty without booke, and where euery Letter standeth: as *b* neere the beginning, *n* about the middest, and *t* toward the end. Nowe if the word, which thou art desirous to finde, begin with *a* then looke in the beginning of this Table, but if with *v* looke towards the end. Againe, if thy word beginne with *ca* looke in the beginning of the letter *c* but if with *cu* then looke toward the end of that letter. And so of all the rest. &c.
>
> (Cawdrey 1970 [1604])

It is understandably difficult for a modern reader to comprehend the need to explain something as now ingrained as ordering items alphabetically, but Cawdrey's labour (along with earlier attempts such as that by Balbus in his 1286 *Summa Grammaticalis*, better known as the *Catholicon*) can give a useful indication of the artificiality of such an arrangement.

The thematic and non-alphabetical thesaurus, by contrast, has a more immediately logical classificatory system, and a longer historical pedigree – although this logic comes at the expense of speed of access when a reader is searching for a particular lexical item. As an alphabetical system is by far the most efficient option when a reader speedily requires a particular structured piece of data about a known and specific word, and as this situation is the prime usage case of a modern dictionary, the alphabet has held an almost unrivalled sway over lexicographical arrangement since Cawdrey's time, despite its disadvantages. Accordingly, most thesauruses also include an alphabetical index, either within their pages or as a separate volume, for ease of lookup.

Nonetheless, given that alphabetical arrangement exists purely for efficiency of access, it is no exaggeration to say that this makes such an ordering intrinsically uninteresting. Alphabetical order tells a reader nothing about the word itself other than its opening configuration of letters, and leads to the alphabetical fragmentation of facts, detaching a dictionary so arranged from any possibility of linking order to meaning.

24.2. The Structure of a Thesaurus

By contrast to this alphabetical ordering, a thesaurus places meaning at the heart of its arrangement. Lexical facts are clustered with other, similar facts, arranged either on the grounds of semantic features, prototypicality effects, usage domains, or thematic harmony. Core to this principle is that, beyond a top-level arrangement devised for pragmatic reasons (see 24.2.1 below), thesauruses are arranged logically and systematically, and it should be apparent to a user how neighbouring entries in a thesaurus are related to each other.

While this system requires an alphabetical index for convenience of lookup, confident and frequent users of a thesaurus such as Roget's become familiar with the system of arrangement, and can find material with some ease. In many historical cases, as outlined below, such thesauruses structures exist either in tandem with, or derived from, an attempt to systematically categorize the world and all of human experience, with the assembly of lexical facts but one part of the overall endeavour. The converse is also true, with lexicographical works such as Roget's and the derived WordNet being popular amongst computer scientists for categorisation systems. The general pattern of arrangement of a thesaurus is outlined below.

24.2.1. Macrostructure

Scratch any thesaurus-maker and you are likely to find not only a lexicographer but also someone ambitious to impose a degree of structure on the apparent disorder of the world around us. To some extent, this is inevitable, since a thesaurus has no predetermined order; one has to choose a starting point, and each category thereafter has to be related in some relatively transparent way to its fellows.

So while a thesaurus is primarily constructed from the bottom upwards, with words and concepts as their main building blocks, they require a high-level structure to be in place during its creation, splitting the world into major classes or categories, under which its other divisions can be placed. It is entirely possible to classify the semantic domain of *Plants*, say, or *Emotions*, in a wholly data-driven fashion, but it is much harder to then continually work upwards to an entirely empirical framework; decisions must be made, often on philosophical, psychological, or (more frequently) pragmatic grounds as to how to represent the large, abstract notions under which everyday life can be subsumed. These top-level classes often include such considerable concepts as Life, the Universe, Matter, and so forth (see further Fischer 2004 for a comparative investigation of these structures). Such choices will often throw light on the intellectual climate and prevailing world-view(s) of the time when the thesaurus was produced. Hüllen (1999), for example, points out the shifting position of God in early classifications: initially usually at the beginning as the omnipotent creator, but increasingly towards the end as part of the social artefact of Religion, though occasionally somewhere in the middle.

Such major divisions form a thesaurus' macrostructure. They are frequently discussed in prefatory matter to a thesaurus (where such is given) or in the lexicographical literature, as their arrangement is necessarily arbitrary, although

ideally logical. At the high level of abstraction these divisions operate, it is difficult to argue for the efficacy or naturalness of any of them, although pragmatic arguments can often be made for each.

The most famous macrostructure is that of Roget, which has 1000 categories split into six major classes, as explained in the discussion of Roget in 24.6.2 below. Roget's macrostructure is shallowly hierarchical, having only a few layers of superordinate categories within which to place the concepts covered by the lexemes it collects, but it nonetheless illustrates the hierarchical framework which is necessary to unite a thesaurus' data-driven microstructure with the abstract macrostructure (the 'top-down' and 'bottom-up' parts of thesaurus-making). By contrast, Tom McArthur's *Longman Lexicon of Contemporary English* (1981) is organized under fourteen major classes (themes), beginning with Life and Living Things, and includes definitions, citations, style labels, illustrations, and grammatical information (McArthur 1986: 147–50; 1998: 149-59).

The most complex thesaurus macrostructure is that of the 2009 *Historical Thesaurus of the Oxford English Dictionary* (HTOED), which was based on the contents of the second edition of the *Oxford English Dictionary* (OED), in addition to other sources. The HTOED system begins with three major divisions – the physical world, the mental world, and the social world – and then proceeds from these into a hierarchical structure which can hold up to twelve nested levels of subcategorisation. This system is necessary to adequately organise the over three-quarters of a million lexemes which this work holds.

Finally, a thesaurus macrostructure can be as idiosyncratic as its compilers wish it to be – for example, P.R. Wilkinson's *Thesaurus of Traditional English Metaphors* bases its classification on a nursery jingle, meaning its highest-level

divisions are *Tinker, Tailor, Soldier, Sailor, Richman,* [sic] *Poorman, Beggarman,* and *Thief,* with Wilkinson adding additional categories of *At Home, At School,* and *At Play*. This results in such startling subcategories such as G.2b "Hostile receptions with mud" and G.2d "Hostile receptions with dogs", contained under *Beggarman* (385-6).

24.2.2. Microstructure

Within a macrostructure, a thesaurus lumps or splits its contents to varying degrees of granularity, as do all other works of lexicography. A distinction can be drawn between those thesauruses which focus on *distinctive* microstructures, attempting to find classes with contents which are recognisably semantically discrete from all others in of similar meaning, and those which have *cumulative* microstructures, wherein the classes assembled contain many words with a relationship of similarity to each other, but are not finely distinguished in their meaning. An example of the former is the HTOED, with 236,000 categories for 797,000 words (an average of 3.4 words per category), while the latter is best exemplified by the 2002 edition of Roget's *Thesaurus*, which has 1,000 categories for "over 300,000 words" (approximately 300 words per category). In the case of a cumulative arrangement, the internal layout of a category list can be in order of frequency of use, alphabetical arrangement, parts of speech, or in a more impressionistic style which depends on the intuition of the compiler.

Each of these two structures has advantages and disadvantages. The primary disadvantage of a distinctive microstructure is the time and energy required to draw apart the subtleties of each word, arrange it into a separate category, and then decide where this category sits in relation to the enormous number of other categories which

are necessitated by this system. For this reason, thesauruses in the distinctive camp often have a complex hierarchical structure to deal with the large number of categories it necessitates. The advantage of this style is that it gives precise and detailed information about each word in turn; similarly, the lack of this precision is often the disadvantage of cumulative microstructures, and tales abound of students misusing Roget by assuming any word in a Roget entry can function as a strict synonym for another. The comparative ease of constructing a cumulative thesaurus, alongside its simpler structure for the user (a consideration mainly important in marketing rather than in actual usability), is the main advantage of this type of structure.

24.3 Particular Types of Thesauruses

The comparative rarity of thesauruses mean that they do not easily form a homogenous grouping. If the modern, synchronic thesaurus such as Roget is taken to be the stereotypical work, then thesauruses which differ from this straightforward model are easily found. They fall into the (non-exclusive) categories below.

24.3.1 Historical Thesauruses

If thesauruses are rare, then historical thesauruses are even rarer. In addition to the usual problems common to the field, historical lexicographers have to engage with issues such as scarcity of evidence, changing world views, and lack of appropriate encyclopaedic knowledge. The basic data usually come from historical dictionaries. It could be argued that creating thesauruses from dictionaries imposes another editorial layer between the lexicographer and the texts; on the other hand, there is little point in repeating work already done by other lexicographers. Problems can arise if the

dictionaries disagree with one another, or the thesaurus-maker disagrees with the dictionary, or if, despite everyone's best efforts, the meaning of an older word is not fully known. Thus, in *A Thesaurus of Old English* (TOE), which classifies the Anglo-Saxon vocabulary surviving in written form from the late seventh century A.D., the editors ended up with a category called simply "Unidentified plants". Plant names are notoriously difficult to interpret — there is no parallel category for the much more readily identifiable animals.

One solution to the problem of changing world-views is to allow the classification to emerge from the words rather than impose a classification upon them. This is the approach taken in Spevack's *Shakespeare Thesaurus*, which derives from an annotated database of Shakespeare's work. He writes that "… a pragmatic cycle of shuffling and filtering and reshuffling of the vocabulary has determined the classification: that is, the names were supplied after the groups began to take shape" (ix). This was also the procedure followed in TOE and the HTOED. All three also set up subcategories when the weight of vocabulary demanded it; to take Spevack's example, the lexicon suggested not only a category of ships, but also categories for various parts of ships, sailors, and navigation. Only when this stage has been reached can one begin to think about the role and importance of seafaring to the Anglo-Saxons, Shakespeare and the Elizabethans, or English speakers over the centuries.

Such folk taxonomies, informed by what Hallig and von Wartburg describe as "naïve realism", guided by "the intelligent average individual's view of the world, based on pre-scientific general concepts made available by language" (cited in Ullman 1962: 255), work well for thesauruses of languages remote from science such as Buck's *Dictionary of Selected Synonyms in the Principal Indo-European Languages* or TOE. Once expert scientific taxonomies become available, and part of

at least some world-views, the classifier may want to take account of both, as John Wilkins did when he was confronted with John Ray's classifications of plants and animals and abandoned his original functional taxonomy of "plants for pleasure", "plants for nourishment" and "plants for medicinal purpose" (Hüllen 1999: 262-3). For HTOED, classifying tens of thousands of words across the entire 1300-year history of English, the expert taxonomy was often the best solution for major scientific sections, though folk categories like Domesticated Animals or Cultivated Plants were included when justified by the data.[i]

Within the macrostructure, historical thesauruses which regard their data as belonging to a single period will usually display synonyms in alphabetically organized lists. Those with a diachronic spread, such as HTOED, will order lists chronologically, or will compromise by including some information about dates of use within an alphabetical list, as the *Scots Thesaurus* does. Headings will usually be in modern English, since it is unrealistic to expect the generality of users to be familiar with older English or Scots. By the same token, older or more obscure words may be omitted from the index. In TOE and HTOED, considerable care was expended on devising headings which would supply more information than is usual in thesauruses by a process of leading back from a specific to a general idea. Thus, in HTOED, one occurrence of the heading "Wedding" leads back through numbered taxonomic stages to "Cake for special occasion" and thence to "Cake", "Dishes and prepared food", and eventually "Food" itself, enabling the reader to create something like a definition.

24.3.2. Synonym Dictionaries

While we have thus far discussed thesauruses as purely semantically arranged, there are also hybrid forms, which are sometimes called synonym dictionaries, containing small semantically-arranged lists of synonyms identified by a headword (eg some synonyms for *pleasant*), with thousands of those small packages then published in alphabetical order. These use what McArthur, discussing alphabetic reference books generally, describes as the "line and blister" model, where the line represents the alphabet and "… each of the blisters represents a special group of synonyms that are best explained together, or a semantic field that should be kept reasonably unified, or a special subject that ought to be covered in depth in one place — despite the alphabet" (1986: 156).

These have the advantage of being somewhat quicker to access than a wholly semantically-arranged work, and are often the dominant sort of thesaurus found in schools, but they also require either extensive cross-referencing (thereby removing their advantage of speed) or to publish the same information many times (with the same list of synonyms for *pleasant* repeated with slight variations under the headwords *enjoyable*, *congenial*, and so on). They include many works with "thesaurus" in their titles, such as the *Collins English Thesaurus* or the *Oxford Paperback Thesaurus*. This style of thesaurus, being a somewhat awkward compromise, although useful for some users, is not discussed here in any detail. Their creation does, however, require a good deal of compromise in balancing the convenience of the alphabet against what we might describe as thematic creep back towards the notional structure of the thesaurus.

24.3.3. Learner's Thesauruses

Thesauruses for learners are popular in two main areas.

The first is where language-teaching works use a thesaurus structure as an intuitive way of learning vocabulary in semantic groups, and large numbers of language-learning textbooks use this system, usually with visual accompaniments for younger readers. In this area, it can be difficult to distinguish the somewhat fuzzy boundary between vocabulary textbooks semantically arranged, which are rarely called ‘thesauruses’ by their compilers or users, and the thesaurus proper.

“Learner’s thesauruses” can also be used to describe an adult-learner-focused thesaurus, which is a standard thesaurus of one of the types above which is annotated, in learner’s dictionary style, with far more metalinguistic annotation and use examples than a traditional thesaurus has. One of the first of these was McArthur’s 1981 *Longman Lexicon* (see 24.2.1 above), and in more recent years this style is represented by the 2008 *Oxford Learner’s Thesaurus*, which provides synonyms and antonyms within an alphabetical list of headwords, alongside usage notes, pronunciations, patterns and usage labels, collocation lists, disambiguation information, and diagrams of scalar synonyms, all with the aim of assisting a learner of English. This thesaurus can be seen as continuing the innovative tradition of learner’s dictionaries in the latter half of the twentieth century.

24.3.4. Domain Thesauruses

The market also offers a handful of technical domain-focused thesauruses on subjects like Art and Architecture, but even these are often synonym dictionaries in disguise. Being limited to a single domain, where meanings are less controversial, expert taxonomies are often available, and polysemy is rarely a problem, they are relatively easy to compile and are often of little interest to the non-expert.

24.4. The Function of a Thesaurus

There are a wide variety of uses which can be found for a thesaurus, in any of the three forms outlined above (distinctive-semantic, cumulative-semantic, and synonym dictionaries). A thesaurus can firstly act as what Haartman and James call an *active dictionary*, one which is "designed to help with encoding tasks, such as the production of a text" (1998 :3). This is perhaps the most common use of a thesaurus as a general reference work, often to assist writers in finding either an alternative term to one they have already used, or a more fitting word for a concept than the one which immediately springs to their mind. Similarly, historical thesauruses can be used to give an air of authenticity to works of historical fiction, amongst others.

Within thesaurus categories, one can also find the range of expressions available to a speaker to encode a concept and thereby see those competing terms that a speaker chose not to use, thereby enabling a literary scholar or a political historian, for example, to discuss word choice by a particular speaker and analyse the subtleties of picking one expression over its opponents.

Similarly, in a comprehensive historical thesaurus, insights can be offered into both language and society. Often the size of a category and its level of detail will indicate the importance of an artefact or concept to a society. TOE, for example, is lexically rich in many aspects of warfare, such as weapons and warriors, showing their role in the literature of the time, or at least in what has come down to us. The chronologically-ordered HTOED can be used to pinpoint areas of lexical growth and decline, for example in technology, foodstuffs, and leisure activities in the modern period. It can also show relationships among words of similar meaning, as when substantial numbers of words of OE origin were displaced by French words after 1066 in domains such as Law and Religion. Sometimes semantic ordering can reveal

connexions between words when there are long gaps in the record: OE *becca* "a pick or mattock" can be linked to OED *beck* in the same meaning, not recorded again until 1875. Those writing dictionary definitions can use a thesaurus to locate a word among its synonyms, which may have changed radically over history. Thus Welsh English *gambo* originally referred to a low, flat cart, but was subsequently extended to other rudimentary or makeshift forms of transport, including an old, dilapidated motor-car, which links it to the HTOED category listing such vehicles. The fact that such a category is already well-established may justify a mention of this meaning in the definition of *gambo* or even recognition of a separate sense.[ii] Wordlists may also throw light on sound symbolism. Examining the HTOED category for *Harsh, discordant sounds*, and the abstract category of *Complaint*, provides fairly convincing evidence for the link between the /gr/ cluster and these concepts (Kay forthcoming).

Much can also be revealed by an in-depth study of a particular field. Wild, for example, traces the development of terminology for 'young person', 'child', and 'baby/infant' and notes how age is increasingly used as a classifier, as in *toddler*, *pre-schooler*, and *teenager*, suggesting "the increased attention paid to children as a section of society" (2010: 298). Alexander and Struan (forthcoming, 2013) discuss the HTOED's *Civilization* category, suggesting five separate metaphorical conceptions of those people discussed as being 'uncivilized' throughout the history of English (namely the categories of wildness, crudeness, foreign-speaking barbarity, incivility, and the state of being significantly Other).

A thesaurus also encodes world views of its compilers. Roget's attitude towards women and sex is notoriously encapsulated in his thesaurus structure and categories, both in how he categorizes parts of the body and in what he omits (the later changes to Roget also map how attitudes differed in the latter twentieth century).

Thesauruses can even be used to predict the future. One of the most original volumes of recent years is Burger's *Wordtree*, described on its title page as "A Transitive Cladistic for Solving Physical and Social problems. The dictionary that analyzes a quarter-million word-listings by their processes, branches them binarily to pinpoint the concepts, thus sequentially tracing causes to their effects, to produce a handbook of physical and social engineering". Having found existing dictionaries "overly humanistic", the editor turned to the language of technology, "an increasingly important part of the mapping of any culture seeking to control its environment". Over a period of twenty-seven years, he collected transitive verbs, analyzed them into binary semantic primitives, and combined them to form a multiply cross-referenced hierarchy of lexical items, where each word is defined by a word from the level above, plus a differentiating component (*Wordtree*: 13-14). This is a book like no other, yet the comment "Each scientific revolution produces a somewhat different grammar and world-view" gives us a clear indication of the use of the thesaurus as an insight into both these revolutions and their associated world-view.

Finally, when a thesaurus is created on the basis of a complete and fixed corpus, as with TOE, a wealth of information is opened up about the nature of not only that language, but also that corpus – the paucity of words in this thesaurus for terms of endearment between two people is not a reflection on the Anglo-Saxon peoples, but rather one on what writings we have of theirs that survive.

24.5. Creating a Thesaurus

Thesaurus creation differs relatively little from the creation of dictionaries in many regards: some raw data, either a corpus, a collection of citations, or another dictionary, is analysed in order to collect a series of lexical facts, the most important

of which for a thesaurus are the word form and its meaning (often these are the only pieces of information included in a thesaurus). The main difference is in the arrangement stage, which in a dictionary is easily done using the alphabet, but in a thesaurus often takes up the bulk of the work. If a macrostructure is already in place, then the body of lexical items involved are then split into the major classes, and then the entirety of such classes are then analysed in turn. The most commonly-used system here is a simple one, involving arranging the large bulk of these words, often in the time-honoured lexicographical paper-slip format, into large groups, then taking each group in turn and creating smaller groups, then taking these smaller groups and categorising them yet more finely, and so on until the desired level of granularity is met. In practice, there is often a lot of cross-reference and cross-pollination of word senses across these working categories. Decisions are also necessarily pragmatic and focused on the data, which makes following a particular theoretical orientation often quite challenging – although it is a notable benefit of such work that it can provide empirical data which can feed into linguistic theories (such as prototype theory in cognitive semantics).

A thesaurus can also be created by adding semantic tags to an existing lexicographic database, as was done with the *Scots Thesaurus*, but in practice this requires a very particular set of skills to be done accurately out of semantic order, and adds another large field of practice to the work of hard-pressed lexicographers, particularly if it is being done as a dictionary is being compiled.

24.6. A Brief History of Thesauruses

Although alphabetically arranged dictionaries now dominate the market, thematically organized thesauruses have a much longer history in the annals of lexicography.[iii]

Hüllen (2009a: 27-28) notes that: “In classical antiquity, and even in the older Chinese, Sanskrit, and Arabic cultures, dictionary-making began with the compilation of lists, for which words were selected according to semantic principles”. These lists might comprise terms for domains such as animals, plants, or family relationships, and were often intended as aids to understanding older texts, such as the Homeric epics. Their purpose was thus didactic, and as they developed they assumed the further purpose of imparting information about the world as well as the terminology needed to discuss it.

Latin texts with marginal or interlinear glosses in Old English (OE) survive from the eighth century A.D. onwards. Over time, these glosses were collected into wordlists, at first related to particular texts, then gathered into independent lists, sometimes alphabetical but often in thematic order. Their primary purpose, as in other parts of Europe, was the teaching of Latin. Favourite topics included the body and its parts, precious stones, medicinal herbs, and natural kinds such as animals, birds, fish, and plants. This practice continued during the Middle English period (1150–1500 A.D.), with glosses also in Anglo-Norman and Old French. Increasing attention was paid to social domains such as the church, civil society, arts and crafts, and the home.

In the fifteenth century, materials for learning vernacular European languages began to appear, stimulated by social changes such as the introduction of printing and increased literacy and mobility among the population. These materials often consisted of multilingual thematic lists, with words from up to eight languages appearing in parallel. English, however, was a low-prestige language at the time, and was rarely included (Hüllen 1999: 105). The Renaissance period also saw the appearance of many new or translated works on technical subjects such as warfare, navigation and horticulture; some of these were accompanied by thematic glossaries.

24.6.1 Organising the world

Stimulated by scientific discovery, interest in classification gained momentum during the seventeenth century, while increased contact with other languages as a result of trade and exploration led to a fascination with the idea of a universal language which might be understood by everyone. An early manifestation of these interests was John Wilkins' *An Essay towards a Real Character, and a Philosophical Language*, published in 1668. This enterprise is based on the assumption that all people perceive the world in the same way, and will therefore be able to communicate common concepts to each other through a set of universal symbols which transcend the limitations of actual languages, thereby returning humankind to happier prelapsarian times. There is no space here to do justice to Wilkins' system,[iv] which presents his universal notions in a structured taxonomy leading from broad general classes such as "vegetative species" to groups of synonyms for individual concepts such as "spending" and "keeping", listed where appropriate with their antonyms. Suffice it to say that, although he aimed beyond a thesaurus of English, Wilkins' work had a profound effect on the subsequent development of thesauruses, and especially on the work of Roget.

Between Wilkins' work and Roget's, there was very little interest in thesauruses. Attention was rather focused on ever larger and more sophisticated alphabetical dictionaries. Such interest as there was in a quasi-onomasiological approach manifested itself in alphabetical dictionaries of synonyms, of which the best known are probably Hester Lynch Piozzi's *British Synonymy* (1794) and George Crabb's *English Synonyms Explained* (1816). Such books often had the purpose of

improving stylistic choice through discussion of the nuances of semantically close words — another development of importance in Roget's work.

24.6.2 Roget's *Thesaurus of English Words and Phrases*

Roget's *Thesaurus of English Words and Phrases* must rank as one of the publishing successes of all time. Writing in 1970, Emblen claims that by then 20 million copies had been sold since the first edition in 1852 (Emblen 1970: 272). By the time of Kendall's biography in 2008, that total had risen to 40 million (Kendall 2008: 1). These figures include the six editions published by Longman between 1852 and 2002, and imprints from other publishers.

Peter Mark Roget lived from 1779–1869 and had a distinguished career as a doctor and scientist before returning after retirement in 1842 to his earlier interest in wordlists. In the *Preface* to the first edition, he reports that " … I had, in the year 1805, completed a classed catalogue of words on a small scale, but on the same principle, and nearly the same form, as the Thesaurus now published" (Davidson 2002: xix). He goes on to say that he had found such lists "of much use to me in literary composition", thus asserting from the outset the main reason for his work's popularity: its usefulness as a resource for those searching for an appropriate word. His comment highlights the key difference between a thematic thesaurus and an alphabetical dictionary. A thesaurus is a productive, encoding device, proceeding from meaning to words expressing that meaning, whereas a dictionary is a receptive, decoding tool, moving from known word forms to their meanings.

In the *Introduction* which follows the *Preface*, Roget explains how he had arrived at a "system of classification of the ideas which are expressible by language … arranging them under such classes and categories as reflection and experience had

taught me would conduct the inquirer most readily and quickly to the object of his search" (Davidson 2002: xxiii). This results in six primary classes: 1. Abstract relations, including Existence, Quantity, Order, Number, Time; 2. Space, including Dimensions, Form, and Motion; 3. Matter, its Properties, and Perception through the five senses; 4. Intellect; 5. Volition; 6. Emotion, Religion, and Morality. Ideally, users of the thesaurus should master this system sufficiently well to be able to home in on the words they need. However, like subsequent editors, Roget was realist enough to appreciate that not all users would have the capacity or commitment to do this, and added an alphabetical index. It is often a sorrow for thesaurus-makers, and an argument against the universality of human conceptual structures, that one person's self-evident system of classification will be largely mysterious to others.

Within the overall structure, Roget offers a flat classification of 1000 categories, subdivided by part of speech, usually with further subdivisions on semantic grounds. Each division or subdivision has a headword of general meaning, followed by lists of what may, by a very generous definition, be regarded as synonyms, but may also include hyponyms, meronyms, and members of the same lexical field. (on these and other sense relations see Murphy, this volume, especially 33.3 and 33.4). As far as possible, the following category contains words of opposite meaning ("correlative terms" - the term antonym was not yet in use), though the practice of laying out the thesaurus in parallel antonymic columns was abandoned from the 1962 edition (Dutch 1962). Roget shows himself well aware of some of problems users and critics may have with his work. He acknowledges the impossibility of completely substitutable synonyms, and, no doubt with experience of synonym dictionaries in mind, the equal impossibility of investigating all the "distinctions to be drawn between words apparently synonymous", intending instead

to “classify and arrange them according to the sense in which they are now used, and which I presume to be already known to the reader” (Davidson 2002: xxvii). Such an assumption has, of course, led to unfortunate results when users have indulged in ill-informed substitution.

24.7 After *Roget*

One sure sign that a product has arrived is when a trade or personal name achieves the status of a common noun, as in “Have you got my Roget?”. Recognition of the merit of the work was not, however, instantaneous. Emblen has some interesting examples of early reviews, and writes: “Most journals and papers that reviewed the *Thesaurus* were reservedly complimentary and somewhat bewildered as to how one would use the thing” (1970: 272). Nevertheless, the popularity of the book grew, and after the crossword puzzle boom hit North America and Britain in the 1920s, it became an indispensable part of any library (Emblen 1970: 278, 281).

American editions of Roget appeared from 1854, with Thomas Y. Crowell and Company taking over as publisher in 1886 and subsequently producing new editions under the title of *Roget’s International Thesaurus* (Emblen 1970: 282). There was some tweaking of Roget’s scheme of classification, for example in Chapman’s classification into fifteen main categories in the fifth edition, on the grounds that Roget’s scheme “…does not coincide with the way most people now apprehend the universe” (*Roget’s International Thesaurus* 1992; quoted in Fischer 2004: 43; see also Hüllen 2009a: 44). A French edition appeared in 1859 with Roget’s approval (Hüllen 2009b: 76; Kendall 208: 266), and there have been versions in German and other European languages (Hüllen 2009b: 60).

From time to time, brave souls make a break for freedom and offer alternatives to Roget's structure, often by choosing a different starting point and reorganising the major classes. Two such were Franz Dornseiff's *Der deutsche Wortschatz nach Sachgruppen* (1933), and Rudolph Hallig and Walther von Wartburg's *Begriffssystem als Grundlage für die Lexikographie* (1952). Dornseiff's classification of German has twenty major classes, beginning with the Inorganic World, followed by Plants, Animals, and Humans, while Hallig and von Wartburg have ten classes in three broad groups: The Universe, Man, and Man and the Universe (Fischer 2004). It should be noted that theirs is a classification of concepts (in French) rather than a classification of the lexicon of a language like Dornseiff or Roget; theoretically, it could be used to display the lexicon of any language. According to Ullman, the scheme caused a good deal of interest when it was presented at the Seventh International Congress of Linguists in 1952, where it was seen mainly as a framework for displaying and comparing different languages or historical periods (1957: 314–5; see also Hüllen 1999: 18–21). There is no record of its being used in its totality or of having much effect on actual thesaurus-making.

A more practical approach is taken in Tom McArthur's *Longman Lexicon of Contemporary English* (1981), the macrostructure of which is outlined in 24.2.1 above. This work, which later influenced the database of the UCREL Semantic Annotation System (USAS), was particularly designed for foreign learners, given the logical use of a thesaurus structure for the learning of vocabulary, and began the later trend for corpus-driven and usage-sensitive learner's thesauruses.

The final set of modern thesauruses which break away from Roget are the historical thesauruses TOE and HTOED, described in 24.3.1 above. These originate

primarily from the work of Professor Michael Samuels, who set up the *Historical Thesaurus of English* project at the University of Glasgow in the 1960s.

24.8. The Way Ahead

No lexicographical field is immune to the disruptive effects of new technology, and this is perhaps most true in the field of thesauruses. The distinction above between a synonym dictionary and a thematic-semantic thesaurus is one which breaks down immediately on entering the electronic arena. A synonym dictionary, which reproduces under given headwords a subset of terms in a semantic thesaurus's microstructure, is only necessary in a printed form; an electronic equivalent would simply be a thesaurus database which dynamically provides to a user a semantic field based on the original search word.

Extending this line of thought, in a time of instant database search results the sole advantage of alphabetical order is entirely lost, and its disadvantages dominate; it places unrelated items next to each other, it loses the opportunity to make connections useful to a user, and it prohibits easy browsing on one particular subject due to its layout. Dictionary data, laid out in a thesaurus structure, becomes the most attractive hybrid, rather than the converse as now. When scholarship and reference works in the digital age require not isolation and fragmentation, but rather union and seamless integration to best serve a user's needs, the strictures of the alphabet become hindrances rather than advantages, and relics of a printed form which is in decline. Without too much hyperbole, it can be easily predicted that the thesaurus, the oldest form of lexicography and one dominant for centuries before the printing revolution, could again become the most dominant form of arrangement in the post-print digital future.

References

Dictionaries

A Dictionary of Selected Synonyms in the Principal Indo-European Languages (1988 [1949]). Ed. Carl Darling Buck. Chicago and London: University of Chicago Press.

A Table Alphabeticall, Conteyning and Teaching the True Writing, and Understanding of Hard Usuall English Words (1997 [1604]). By Robert Cawdrey. Ed. Raymond E. Siemens, Toronto: University of Toronto Library.

Begriffssystem als Grundlage für die Lexikographie: Versuch eines Ordnungsschemas (1952). By Rudolph Hallig and Walther von Wartburg. Berlin: Akademie-Verlag.

Der deutsche Wortschatz nach Sachgruppen (2004 [1933]). By Franz Dornseiff. Berlin and New York: de Gruyter,

An Essay towards a Real Character, and a Philosophical Language (1968 [1668]). By John Wilkins. Menston: Scolar Press.

Historical Thesaurus of the Oxford English Dictionary (2009). Ed. Christian Kay, Jane Roberts, Michael Samuels, and Irené Wotherspoon. Oxford: Oxford University Press.

The Longman Lexicon of Contemporary English (1981). Ed. Tom McArthur. Harlow: Longman.

The Oxford English Dictionary (1884-1928). Ed. Sir James A. H. Murray, Henry Bradley, Sir William A. Craigie and Charles T. Onions. *Supplement and Bibliography* (1933). *Supplement* (1972-1986); ed. Robert W. Burchfield. 2nd. ed., (1989); ed. John A. Simpson and Edmund S. C. Weiner. *Additions Series*, (1993-7); ed. John A. Simpson, Edmund S. C. Weiner and Michael Proffitt. 3rd. ed. (in progress) *OED Online* (March 2000-), ed. John A. Simpson, www.oed.com [= *OED*]. Oxford: Oxford University Press.

The Oxford Learner's Thesaurus (2008). Ed. Diana Lea. Oxford: Oxford University Press.

Roget's International Thesaurus, 5th ed.(1992). Ed. Robert L. Chapman. New York: HarperCollins.

Roget's Thesaurus of English Words and Phrases. (1962 [1852]). Ed. Robert A. Dutch. Harlow: Longman.

Roget's Thesaurus of English Words and Phrases: *150th anniversary edition* (2002 [1852]). Ed. George W. Davidson. London: Penguin. (check refs.)

The Scots Thesaurus (1990). Ed. Iseabail Macleod with Pauline Cairns, Caroline Macafee, and Ruth Martin. Aberdeen: Aberdeen University Press.

A Shakespeare Thesaurus (1993). Ed. Marvin Spevack. Hildesheim, Zurich, and New York: Olms.

A Thesaurus of Old English (1995). Ed. Jane Roberts and Christian Kay with Lynne Grundy. London: King's College. 2nd ed., 2000. Amsterdam: Rodopi.

The Wordtree (1984). Ed. Henry G. Burger. Merriam, Kansas: Wordtree.

Thesaurus of Traditional English Metaphors, 2nd ed. (2002). Ed. P. R. Wilkinson. London and New York: Routledge.

Secondary sources

Alexander, Marc and Andrew Struan (forthcoming, 2013). "'In countries so unciviliz'd as those?': Notions of Civility in the British Experience of the World", in Martin Farr and Xavier Guégan (eds.) *Experiencing Imperialism*. London: Palgrave Macmillan.

Emblen, D. L. (1970). *Peter Mark Roget, The Word and the Man*. London: Longman.

Fischer, Andreas (2004). 'The notional structure of thesauruses', in Christian J. Kay and Jeremy J. Smith (eds), *Categorization in the History of English*. Amsterdam and Philadelphia: Benjamins, 41-58.

Haartman, R. R. K. and James, Gregory (1998). *Dictionary of Lexicography*. London: Routledge.

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

Hüllen, Werner (2004). *A History of* Roget's Thesaurus: *Origins, development, and design*. Oxford: Oxford University Press.

Hüllen, Werner (2009a). 'Dictionaries of Synonyms and Thesauruses', in A. P. Cowie (ed.), *The Oxford History of English Lexicography, vol. II: Specialized Dictionaries*. Oxford: Clarendon Press, 25-46.

Hüllen, Werner (2009b). *Networks and Knowledge in* Roget's Thesaurus *from ancient to medieval*. Oxford: Oxford University Press.

Kay, Christian (2010). 'Classification: Principles and Practice', in Michael Adams (ed.), *Cunning Passages, Contrived Corridors: Unexpected essays in the history of lexicography*. Monza: Polimetrica, 255-270.

Kay, Christian (2012). 'The Historical Thesaurus of the OED as a Research Tool', in Kathryn Allan and Justyna Robinson (eds), *Current Methods in Historical Semantics*. Berlin: Mouton de Gruyter, 41-58.

Kay, Christian (forthcoming). 'Some Interesting Sounds in the *Historical Thesaurus of the Oxford English Dictionary*'. *SELIM proceedings.* Convergent Approaches to Mediaeval English Language and Literature.

Kendall, Joshua (2008). *The Man who made Lists: Love, death, madness, and the creation of* Roget's Thesaurus. New York: G. P. Putnam's Sons.

McArthur, Tom (1986). *Worlds of Reference.* Cambridge: Cambridge University Press.

McArthur, Tom (1998). *Living Words: Language, lexicography and the knowledge revolution.* Exeter: University of Exeter Press.

O'Hare, Cerwyss (2004). "Folk Classification in the HTE Plants Category", in Christian J. Kay and Jeremy J. Smith (eds), *Categorization in the History of English*. Amsterdam and Philadelphia: Benjamins, 179-191.

Ullmann, Stephen (1957). *The Principles of Semantics*, 2nd ed. Oxford: Basil Blackwell.

Ullmann, Stephen (1962). *Semantics: An introduction to the science of meaning*. Oxford: Basil Blackwell.

Wild, Kate (2010). 'Angelets, Trudgeons, and Bratlings: The lexicalization of childhood in the *Historical Thesaurus of the Oxford English Dictionary*', in Michael Adams (ed.), *Cunning Passages, Contrived Corridors: Unexpected essays in the history of lexicography*. Monza: Polimetrica, 290-308.

---

[i] For a discussion of the issues involved in classifying plants, see O'Hare (2004). Further details about HTOED can be found in its *Introduction* and in Kay (2010; 2012).

[ii] We are indebted to Andrew Ball of the OED for this example.
[iii] This section draws heavily on Hüllen (1999), to which the reader is referred for a much more detailed account of the early history of thesauruses.
[iv] A full description and analysis is given in Hüllen (1999), chapter 8.