12

Lexicography

The construction of dictionaries and thesauruses

Pam Peters

12.1 Introduction

Ferdinand de Saussure’s (1916) model for the words of a lexicon – consisting of the signifier (word form) and the signified (word meaning) – set the foundations of modern linguistics. These complementary perspectives also correspond with the two conventional formats for representing a lexicon: the dictionary and the thesaurus. The dictionary is an inventory of the word forms of a language, the thesaurus a complex semantic structure designed to account for all the meanings embedded in the lexicon. In lexicographic terms, the dictionary format is semasiological, and that of the thesaurus onomasiological. Those not entirely transparent words may be explained by the Greek roots they embody: the term semasiological is derived from sema the word for ‘sign’, since dictionaries are constructed out of the regular signifiers of the language; while onomasiological is derived from onoma ‘name’, since thesauruses provide labels for the concepts and meanings (the signifieds) of the lexicon within a multilayered semantic structure.

Elementary dictionaries and pioneering thesauruses were published well before de Saussure, the earliest English examples both dating from the seventeenth century. The first small monolingual English dictionary was Robert Cawdrey’s (1604) Table Alphabeticall … of Hard Usuall English Words; while the English thesaurus was pioneered by John Wilkins (1668) in his bulky Essay Towards a Real Character and a Philosophical Language, using special symbols to designate words and their meanings within a large conceptual hierarchy. These early models were enlarged in the print medium in the centuries following, in response to socio-cultural needs. Each innovation prompted changes in the construction of lexical information within the two types of language reference, as discussed in §§12.2 and 12.3. Since the turn of the millennium, both dictionary and thesaurus have continued to evolve as lexical references in the electronic medium, as tools for natural language processing (NLP) and artificial intelligence (AI), and as interfaces within information systems. Computerized lexica serve an ever-widening spectrum of human and machine readers. In the process, both dictionaries and thesauruses have extended their scope and added features of the other.

In principle the dictionary and the thesaurus offer language-users alternative ways of accessing the vocabulary, via the known word forms or the abstract concepts whose...
representation they contribute to. However the terms *dictionary* and *thesaurus* were and still are polysemous, so that printed works and computer software bearing those names do not necessarily conform to the archetypal semasiological or onomasiological construction noted by twenty-first century lexicographers. The first monolingual English dictionaries to be titled as such, embrace the general English lexicon, and construct more than minimal alphabetical entries about each word were eighteenth-century products. They demonstrated the utility of the dictionary’s alphabetical construction in accessing general and/or specialized words of the language – hence its widespread use in the titles of reference books of itemized information in any field of knowledge, e.g. *Dictionary of Dreams*, *Dictionary of Sport and Exercise Science*. For thesauruses there is no conventional set of conceptual structures. Their construction varies with the publication, though each is designed as a means of accessing the common lexicon, to help those reaching for words lost from short-term memory, or wanting to explore sets with similar or related meanings. Yet thesaurus structures do not seem to be as intuitive as their authors intended, and most thesauruses provide an alphabetical index as a key to their internal construction (see §12.3.3).

This chapter traces the development and increasing convergence of dictionaries and thesauruses in English since the early modern era. From their different practical and theoretical starting points, both have evolved in structure and content in response to the pedagogical and socio-cultural needs of the times. Commercial imperatives have no doubt always impacted on their construction, as well as recent quantum leaps in information technology. Whether lexicography can be said to have developed a body of theory is an open question (Wierzbicka 1985), amid debate about its independence as a discipline. Lexicographers divide over its debt to lexicology and (applied) linguistics, though both theorists and practitioners agree on the importance of foregrounding the dictionary’s functions and users (Bergenholz and Tarp 2003; Rundell 2012). Some suggest it would be more appropriate to speak of the *science* of lexicography (Bogaards 2010), now that the body of empirical research on dictionary use has been effectively indexed (Welker 2010). With the integration of semasiological and onomasiological aspects of the lexicon in twenty-first century information technology (see §12.4), any emerging theory would need to encompass both dictionaries and thesauruses.

### 12.2 Dictionaries

#### 12.2.1 The construction (macrostructure/microstructure) of dictionaries and their evidential base

Since the late twentieth century, the construction of dictionaries has been seen in terms of two complementary structures: macrostructure (the itemized inventory of words included), and microstructure (the structural template in which information on each word is provided at its entry). Over the course of centuries (see §§12.2.2, 12.2.3) both the overall macrostructure and the microstructure of entries in English dictionaries has steadily enlarged, in response to the needs of the times and different kinds of user. The dictionaries published show waves of interest in general and specialized vocabulary, in common nouns and proper names, in providing purely lexical information and adding encyclopaedic information about the items listed. Through all these shifts in focus and content, the alphabetical macrostructure has proved accommodating and provided easy access to individual items, despite the arbitrary juxtapositions of items within it. The alphabetical macrostructure has also supported alternative approaches to the lemmatization of words:
• clustering morphologically related words, e.g. fossil, fossiliferous, fossilize together in a single entry with the derived forms as secondary headwords or runons, as in the Oxford dictionary tradition;
• splitting morphologically related words over successive individual entries, so that each one is presented in a separate entry, as in Merriam–Webster’s dictionaries.

These different approaches to the construction of the macrostructure are informally referred to as ‘lumping’ and ‘splitting’. They are extremes in terms of macrostructural design, and modern dictionaries usually find a compromise in between. The dictionary’s policy on lumping and splitting affects the microstructure of individual entries, which are consistently larger or smaller accordingly. That apart, the microstructure of monolingual dictionaries has steadily evolved from the minimalism of the earliest monolingual English glossaries (§12.2.2), to the greatly enlarged entries of Johnson’s dictionary (1755) and the first edition of the OED (1884–1928), discussed in §12.2.3.

The interplay between macrostructure and microstructure is also reflected in the dictionary’s perspective on the lexicon: synchronic or diachronic. The synchronic dictionary lists only words deemed current for the contemporary reader (as in Cawdrey’s list of ‘usuall’ words), whereas the macrostructure of the diachronic dictionary includes obsolete words known from the earlier literature of the language. Complementing this, the diachronic dictionary orders word senses/definitions in line with their historical development (as shown by the citational record), and showing whether their usage continues or has ceased. In the microstructure of synchronic dictionary, the order of word senses is not usually historical, but according to their notional generality or specificity – hence the practice of defining undertaker first as ‘one who undertakes’ and then as ‘one who manages funeral arrangements for another party’. An alternative principle for ordering senses is to prioritize high-frequency over lower frequency usages – a system that can be supported by referring to computer corpora, and applied in recent pedagogical dictionaries.

Between the macrostructure and microstructure, lexicographers since Nielsen (2003) have noted as mediostructure the various intermediate structures of the dictionary, which serve as a bridge between the alphabetical macrostructure, and the details of the microstructure in individual entries. The commonest form of mediostructure is the cross-referencing signalled within the microstructure that can take readers to an alternative word, name, spelling or form of the headword, as in:

throve a past tense of thrive

Cross-referencing is the sole function of ‘blind’ entries, used as place-holders for words which are detailed in other entries:

Ulysses butterfly ⇒ mountain blue butterfly

While print dictionaries have long included cross-referencing, many other kinds of mediostructure can be built into online dictionaries. They can include links to supplementary lexical information within the website, such as hierarchical diagrams that show sense relations among word sets (e.g. hyponymy, meronymy), the grammatical paradigms in which word forms vary (e.g. irregular verbs), and computerized concordances extracted from corpora showing the headword as a keyword in context. Some lexicographers (Nielsen 2003) also regard the external links from an online dictionary to resources held elsewhere
on the internet (e.g. formal standards), as part of the dictionary’s mediostructure. These extensions to the information provided to the user of online dictionaries underscore their increasing role as interfaces to the encyclopaedic world of the internet: see further §12.2.4 and §12.4.

Since Johnson’s dictionary, citational evidence from outside sources has been provided within large monolingual dictionaries as a way of validating the individual senses of words, and illustrating their grammatical use in sentences and types of discourse. In the diachronic Oxford English Dictionary (OED), this citational evidence became the backbone of the word’s history, from which its senses over the course of time were induced. It was the work of many hands (see OED’s Historical Introduction, Section VII), though nowadays largely superseded by the availability of computerized databases of printed text (such as the British National Corpus) or data extracted from the internet. Linguistic corpora have been used by the major dictionary publishing houses since the 1980s, starting with Collins publishers’ ‘Bank of English’, and the Oxford Text Archive. Used appropriately, these sources provide both frequency information, and lively citations from different kinds of texts to illustrate more and less formal styles of usage. The polysemy and collocational behaviour of very common words is more easily analysed on concordances derived from corpus data, as demonstrated by Sinclair (1991) with the many uses of of. The ready availability of citational evidence from digitized sources has freed dictionary-makers from their dependence on pre-existing works, and standing on the shoulders of their predecessors, while avoiding plagiarism. However this does not lessen the need for lexicographic scholarship in interpreting lexical raw material.

12.2.2 Origins of the English dictionary as a pedagogical tool

The earliest recorded use of the word dictionary (dictionarius) in England (c.1225) is associated with the teaching of Latin in medieval schools. At that stage it referred to a list of Latin words, and only centuries later did it become the regular name for the familiar monolingual or bilingual dictionary. Even in the mid-sixteenth century, an English–Latin teaching manual titled A Shorte Dictionarie for Yong Begynners (John Withals 1553) presents words and their explanations in topical chapters like lesson plans, rather than alphabetically. The first monolingual alphabetical glossaries of English were published in the late sixteenth century as supplements to Tudor teaching manuals, such as those of Robert Mulcaster in The Elementarie (1582), and Edmund Coote in The English Scholemaister (1596). Each of them extended his predecessor’s inventory with additional ‘hard’ words borrowed into English from Latin, Greek and Hebrew through Renaissance scholarship. Robert Cawdrey’s Table Alphabeticall (1604) updated that of Coote (Stein 2010) to comprise around 2,500 words. It is generally recognized as the first English dictionary, despite its title and large amounts of legacy content (Landau 2001), no doubt because it was published as an independent reference work. Cawdrey is the first to explicitly envisage use of the dictionary outside the schoolroom, suggesting in his subtitle that the book would benefit ‘ladies, gentilewomen or any other unskilfull persons’.

Cawdrey’s work exemplifies aspects of later lexicographical practice, though not consistently implemented. His alphabetization of headwords is sometimes slightly erratic; plural headwords e.g. affaires, records, are mixed in with the standard singular forms, and inflected forms of verbs used as adjectives, thus not fully lemmatized to the base forms (Stein 2010). The source language for borrowed words is indicated by means of an abbreviation, e.g. [g] or [gr] for Greek, and the paragraph sign for French. Cawdrey’s
wordlist is fully glossed, but the explanations are a mix of synonyms used like the translation equivalents of earlier bilingual dictionaries, and definitional paraphrases. Some definitions are simply generic, as when artichock is glossed as a type of ‘herbe’, and beagle as a type of ‘hound’. Alternative meanings are run on within the line of paraphrase, separated by commas, or by or, without marking distinct grammatical roles:

adocate, a spokesman, attourney, or man of law, plead

Occasionally the grammar of the definition is at odds with that of the headword itself, as in

docilitie, ease to be taught

Grammar is clearly an underdeveloped aspect of the Table Alphabeticall. Yet the language of Cawdrey’s definitions is usually clear, and not resolutely impersonal as in later dictionary practice (cf. §12.2.4). An authorial I/we appears in defining meanings in space and time (Stein 2010), as in:

hemisphere, halfe of the compasse of heauen, that we see

moderne, of our time

Later, extended editions of Cawdrey’s wordlist in 1609, 1613 and 1617 are testimony to its popularity.

The ‘hard words’ tradition in monolingual English lexicography grew steadily with seventeenth century dictionaries that contained countless foreign loanwords annexed into English through its expanding functions as the national language (Peters 2012). Many certainly belonged to literary style and formal registers of usage rather than the speech-based styles advocated by the Royal Society (Gordon 1966: 126–8). Thomas Blount’s Glossographia (1656) included about 9,000 borrowed words (with discursive and sometimes far-fetched etymologies), and volumes of technical terms from the arts and sciences, with explanatory information. Other dictionaries in the ‘hard words’ tradition soon followed: Edmund Phillips’s New World of English Words (1658) with 11,000 words, and Elisha Coles’s An English Dictionary (1676), whose headword list expanded to 25,000 with the addition of dialect words, and underworld slang – which he suggests ‘may perchance save your throat from being cut, or… your pocket from being pickt’ (Starnes and Noyes 1946: 63). To make room for these extra words, Coles compressed his definitions and most are just synonyms.

Despite their expanding inventories, seventeenth century ‘hard words’ dictionaries do not show advances in English lexicographic practice. In the same period, the frontiers of monolingual lexicography were being advanced within European academies, notably the Italian Accademia della Crusca and the Académie Française, as a means of establishing the status and scope of the national ‘vernacular’ (Tucker 1961: 58, 90). The lack of such an academy in England meant there was no concerted effort to codify the common English language, and English lexicography was still limited by its origins in bilingual dictionaries and compendiums of specialized vocabulary.
12.2.3 Key developments in the construction of monolingual English dictionaries

The watershed for monolingual English lexicography was the turn of the eighteenth century, with the publication of dictionaries which embraced both everyday and less common words within the lexicon, thus a comprehensive listing of the general vocabulary of English. The first dictionary to do so was *A New English Dictionary*, published in 1702 by ‘JK’, 6 with a large inventory of 28,000 words probably extracted from pre-existing spelling lists and bilingual dictionaries of Latin and French (Osselton 2011). However JK’s glossing was rudimentary, and only in the second edition (1713) was each word given a simple definition. Compare:

\[(1702) \quad \text{To Gaggle like a goose}\]
\[(1713) \quad \text{To Gaggle, to cry like a goose}\]

JK’s dictionary is far from discriminating alternative senses, or providing information on word grammar or etymology. Those now typical features of monolingual general lexicography are first seen in Nathan Bailey’s *Universal Etymological English Dictionary* (1721). Amid its inventory of about 40,000 words (Starnes and Noyes 1946: 100), *about* and *above* appear among countless words drawn from specialized fields such as law, medicine and horsemanship. All are carefully defined, with alternative senses treated according to their grammatical roles, and given specific or generic etymologies, except for very common words. Bailey’s single-volume octavo publication significantly advanced the macro- and microstructure of the English monolingual dictionary.

Samuel Johnson’s *Dictionary of the English Language* (1755) eclipsed all before it, by its sheer size (two folio volumes), and the enormous amount of scholarship brought to bear on documenting individual words and their senses from the canon of English literature. Its headword list consisted of the common English vocabulary, with a small proportion of cant, i.e. slang words, but not proper nouns. Individual entries were all given etymologies, whether they were foreign loans or constructed out of the common stock of English morphemes. Each word’s grammar was noted, in terms of its word class, and different senses were explained and enumerated, and often illustrated with citations from the English literary tradition, as in the following:

**BARGE** n.s. [from *bargei* Dut., from *barga* Low Latin]

1. A boat for pleasure

   \[\ldots \text{the barge she sat in, like a burnish’d throne}\]

   *Burnt on the water* Shakesp. Antony and Cleopatra

2. A boat for burden

Johnson’s dictionary was reprinted with little editorial change for seventy years (Reddick 1990: 171,175–6), and it set the benchmark for English lexicography well into the Victorian era. It appeared at an opportune moment to be the supreme lexical reference for the national language of Britain, sampling the ‘flower’ of the English literary tradition. Its citational
material from Shakespeare onwards served to document thousands of words and senses which were notionally current in mid-eighteenth century England, yielding a more or less synchronic account of the lexicon. But its perspective seemed decadent to nineteenth century scholars: its microstructure deficient in the treatment of the history of word meanings – too much geared to making English a world language (Willinsky 1994: 16–17) – while its macrostructure lacked coverage of foreign loanwords encountered abroad (Ogilvie 2013: 61–8). At the same time the earlier history of English was opening up through research associated with the English Philological Society, which helped to mobilize the New English Dictionary project that became the Oxford English Dictionary (OED). Like Johnson, it aimed to provide citations for the senses of all common words, but presented them as a chronology of the word’s growth and/or narrowing of usage over the centuries. It thus projects a longitudinal/diachronic view of the English lexicon, with both current and obsolete headwords listed. The OED noted etymologies for all words, as well as changes in their spelling over time, and modern pronunciations in an IPA-like script. While it documented changes in word usage through citations, the OED’s founding principle was to avoid making judgements about them as good or bad usage – letting the citations speak for themselves (Willinsky 1994: 17–18). This descriptive and inclusive stance was resolutely maintained by the chief editor (James Murray), despite some criticism of it (Ogilvie 2013: 54).

The OED was unique among English dictionaries in its diachronic scope, its regular use of citations, and its sheer size (twelve volumes published successively from 1884 to 1928). It stands apart from the mainstream of modern English lexicography, which flows through the single-volume synchronic dictionary that provides pronunciations, orthographic and grammatical information, and definitions for all the enumerated senses of each word. Citations (invented or extracted from other sources) and summary etymologies may or may not be offered, depending on the size of the publication. Editorial advice on contentious usage may be added, apart from hidden judgements expressed through the exclusion of ‘unacceptable’ usages. Thus the details contained in the dictionary microstructure vary considerably, depending on the intended users and functions, as discussed in the following section. The makeup of the macrostructure varies also, in the numbers and types of headwords out of which it is constructed, as discussed in §12.2.1.

12.2.4 Functions and users of dictionaries, general and specialized

Dictionaries have long been associated with language learning, as is clear from their earliest uses in medieval and Tudor schools (§12.2.2). The early bilingual Latin–English and English–Latin glossaries were intended to support the teaching and learning of Europe’s first language of scholarship in Britain; while the first monolingual dictionaries of the early seventeenth century were to help adult readers with the anglicized loanwords (‘hard words’) appearing in contemporary texts. The native-English-speaker was the target user of all these works, and their knowledge of English grammar and idiom was taken for granted. As detailed definitions were added into the microstructure of eighteenth-century dictionaries, their rather formal non-propositional language – ‘dictionaryese’ – was something readers were expected to cope with. Native-English-speakers were the assumed readers of monolingual English dictionaries until the later twentieth century, when the needs of foreign learners began to be recognized by major English dictionary publishers, in what are now known as ‘learners dictionaries’.

The English learners dictionary has evolved quickly since the 1960s (Marello 1998), making pedagogical lexicography a driving force in adapting the construction and content
of dictionaries to the needs of dictionary users (Tarp 2009). Its impact can be seen in the dropping of low frequency words from the macrostructure, and greatly increased articulation of the linguistic properties of words in the microstructure. Notable additions are in the semantic keywords, in flexible approaches to definitions and ensuring their accessibility, in extensive use of examples to show natural usage of the word, and some indication of its relative frequency in written and spoken English. All these can be found in the Longman Dictionary of Contemporary English (LDOCE, third edition 1995) for a long entry on the adjective hot with twenty-seven different senses and distinction collocations, grouped with definitions and examples under the following set of keywords:

1. high temperature 2. hot taste 3. difficult to deal with 4. angry 5. popular 6. following closely 7. other senses

Frequency information given at the start (SL, W2) shows that hot is used more often in spoken than written English, based on its ranking within the first 1,000 words in spoken corpus data, and the first 2,000 words in written data. LDOCE uses a controlled vocabulary of around 2,000 words/lemmas in its definitions to ensure they are accessible to those with limited knowledge of English. It makes some use of the ‘sentence definitions’ pioneered in the Collins Cobuild learners dictionaries, as in:

If you are hot, your body is at an unpleasingly high temperature

(Collins Cobuild Essential English Dictionary 1988)

Lexicographic innovations like these for second language learners have yet to be matched in the ordinary dictionaries for native-speaking school students. Recent research on Dutch age-graded dictionaries shows they tend to overestimate the level of language competence required for their use (Schryver and Prinsloo 2011). Continual testing of their microstructure is needed to ensure learners dictionaries fulfil their pedagogical function as effectively as possible.

Specialized dictionaries of technical terms are published in ever-increasing numbers to meet the needs of learners and workers in a particular field, with the continuously expanding terminologies in areas such as chemistry, environmental science, information science, medicine (Ayto 1999). Amid technological advances, the dictionary macrostructure quickly gets out of date, apart from the commercial pressure to keep expanding its market/readership. So the original Mosby’s Medical Dictionary is now Mosby’s Medical, Nursing and Allied Health Dictionary, appearing at ever-closer intervals of time (2002, 2006, 2009, 2011). The current headword list still includes countless medical terms drawn from Latin and Greek, e.g.

metacarpal phalanx … the hands and fingers, particularly phalanges that articulate with carpal bones

as well as those constructed out of common English for occupational syndromes, their diagnosis and nursing care:

metal fume fever … an occupational disorder caused by the inhalation of fumes of metallic oxides and characterized by symptoms similar to those of influenza …. Access to fresh air and treatment of the symptoms usually alleviate the condition

(Mosby’s Dictionary 2011: 1085–6)
The second example illustrates the encyclopaedic element often built into the microstructure of specialized dictionaries, to inform those whose training crosses over from one discipline to another in the larger domain of medicine and health care. It reflects the dictionary’s auxiliary function as repository of knowledge (Bergenholz and Tarp 2003), not just a terminological standard, or strictly linguistic reference. This encyclopaedic dimension has opened up especially for online specialized dictionaries, whose pages can be linked to illustrations, audio files and other multimedia enhancements of their microstructure, as well as external sources of information.⁹

Encyclopaedic features in the dictionary’s microstructure can also be found in certain types of bilingual dictionaries, especially those designed to conserve and maintain endangered languages and sometimes also to revitalize their use (Ogilvie 2011: 394–99). Their socio-cultural function is often visible in the presence of terms relating to traditional crafts in the macrostructure, accompanied by drawings and information in the microstructure to explain their place in cultural practice. Thus the entry for the *bilirra* (‘yellow mangrove tree’) in the *Kayardild Dictionary and Thesaurus* (Evans 1992) includes a line drawing of the tree’s stand of roots above the waterline, and notes that they are used as paddles. Conservation dictionaries of endangered languages often involve partnerships between local people and linguist-lexicographers, bringing lexicographical skills and techniques to bear on endangered cultures (Ogilvie 2011: 393–5). Such dictionaries are important in terms of adding to the documentation of the world’s stock of languages, as well as affirming the native speakers’ language and giving it some status on the world scene. Their role is analogous to that of any national dictionary, including Johnson’s 1755 *Dictionary of the English Language*, which fulfilled the key cultural function of codifying eighteenth-century English with a comprehensive record of the lexicon and extensive illustration of the English literary tradition (see §12.2.3).

12.3 Thesauruses

12.3.1 Origins of the English thesaurus and precursors to Roget

The term *thesaurus* is neo-Latin, based on the classical Greek word for ‘treasury’, i.e. storehouse for precious objects. But among its various applications in English it is rarely found in its classical sense, and its application by lexicographers to an onomasiological lexical reference was not its earliest English use. From the sixteenth century to the eighteenth the term *thesaurus* appeared in the title of semasiological dictionaries, such as the alphabetically arranged bilingual *Thesaurus Linguae Romanae et Britannicae* (Cooper 1565). The term’s application to a conceptually ordered presentation of the lexicon began with Roget’s *Thesaurus* (1852), yet it continued to be applied to alphabetically ordered language references, as in the *American Thesaurus of Slang* (1942). Since the later twentieth century, the term *thesaurus* has been extended to structured information systems created for libraries and information technology (see §12.3.5). Lexicographers themselves use *thesaurus* in more than one sense, so that alongside its Rogetian sense of being a concept-driven construction of the lexicon at large, *thesaurus* may refer to a lexical index to a ‘closed corpus’ (Ilson 2010), i.e. a finite language database such as the *Thesaurus of Old English*, or a published dictionary, as in the case of the *Historical Thesaurus of the Oxford English Dictionary*. The term *thesaurus* is thus polysemous, though its use to refer to a strictly onomasiological description of the English language has been trademarked in the UK as *Roget’s Thesaurus*. 
The first truly onomasiological work with reference to the English language was produced in the seventeenth century by members of the British Royal Society, led by John Wilkins (1614–1672). Published as An Essay Towards a Real Character and a Philosophical Language (1688), it included exploratory discussions on the value of creating a symbolic system to express universal human thoughts, a pervasive theme among contemporary European philosophers such as Descartes and Leibniz. To achieve this goal, Wilkins devised a system of ‘real’ characters (non-alphabetic shapes) that were intended to express human thought directly, without recourse to the words of any individual language. Instead the system should name ‘things and notions’, classified into structured semantic sets so as to function within conceptual hierarchies like those of the modern thesaurus. The overarching conceptual systems were set out in detailed ‘philosophical tables’ (Clauss 1982: 540), drawn up by Wilkins’s colleague William Lloyd. The words/names in the tables were painstakingly cross-referenced to the ‘alphabetical dictionary’, the largest and last section of the Essay (Considine 2008: 298). The Essay’s dictionary presents the inventory of 11,500 English words identified for the things/notions of the theoretical universal language, excluding words which named anything that was specific to place or time (Clauss 1982: 544).

Despite its originality, Wilkins’s monumental Essay was not particularly well received outside the Royal Society (Clauss 1982: 542–6), and its grand onomasiological design was eclipsed for nearly two centuries by other types of dictionary constructed on a more limited semantic basis, i.e. covering a set of semantic relations or a specific semantic field. One was the eighteenth-century ‘topical dictionary’, whose authors were the pioneers of terminography in applied sciences and contemporary areas of knowledge (e.g. art, health). Their focus on a particular field of endeavour arguably gave their work a conceptual/onomasiological basis, even within an alphabetical macrostructure (Hüllen 1999). Another specialized type of dictionary – dictionaries of synonyms – can also be seen as maintaining the onomasiological approach in the long interval between Wilkins’s Essay and Roget’s Thesaurus. Although the key words to each synonym cluster are alphabetically listed, the microstructure for each entry presents a semantic/conceptual analysis of the cluster and the finer sense distinctions among them. The model for English synonym dictionaries was the work of Abbé Girard (1718), La Justesse de la langue françoise ou les différentes significations des mots qui passent pour synonymes, which was translated and adapted to the English lexicon by John Trusler (1766). An independent and better known perspective on English synonyms was published later in the century by Hester Piozzi, who had in a previous marriage (as Mrs Thrale) been an acquaintance of Samuel Johnson. Her dictionary’s formidable title: British Synonymy: Or an Attempt to Regulate the Choice of Words in Familiar Conversation (1794) resonates with the standardizing impulse of the eighteenth century. Yet inside the covers it was evidently designed as a contribution to the lexical education of middle-class women, to help them converse elegantly, and ‘intended chiefly for the parlour window’ rather than the library shelf. It was lively and well received – republished twice (Gove 1984). The writers of synonym dictionaries after Piozzi (notably William Taylor, English Synonymes Discriminated (1813); George Crabb, English Synonymes Explained (1816)) reverted to more discursive treatments of sense relations traceable back to Girard (Gove 1984). These various lexicographic endeavours suggest growing interest in semantic fields within the English lexicon, while the alphabetical macrostructure remained the default for presenting lexical information.
12.3.2 The construction of Roget’s Thesaurus

Roget’s *Thesaurus* (1852) of the English lexicon was a quantum leap in providing an onomasiological account of a European language. Though Roget can be seen as redeveloping the resources of synonym dictionaries, the conceptual structure he devised was original, an analytical system he had begun decades earlier (in 1805), and developed in the background to his professional work as a doctor, scientist and as secretary of the British Royal Society from 1827 to 1848. His interests were also philosophical – like Wilkins, he engaged with Leibniz’s philosophy and questions of how human thought was represented in language. That apart, he had a practical concern with finding the right word to express an idea, and for many users of his *Thesaurus*, it was effectively a ‘word-finder’. The macrostructure of Roget’s *Thesaurus* is a set of hierarchically ordered concepts, articulated vertically down from just six at the highest level:

- abstract relations
- space
- matter
- intellect
- volition
- sentient and moral powers.

Beneath these ‘superconcepts’, he developed a structure of 1,000 concepts and subconcepts, so that the words labelling them are in hyponymic relations with those above and below. Some of the nodes at the second and third levels down had to be labelled by means of phrasal descriptors, e.g. ‘power in operation’, ‘indirect power’ (both under CAUSATION) because there was no single word for the concept. These phrasal descriptors show gaps in the hierarchies of nomenclature for ordinary words, unlike scientific terminology where there is hyponymic nomenclature at every level within its taxonomic systems.

At the fourth level of the *Thesaurus* hierarchy, Roget provided two sets of words that express a given concept: one containing analogous words, the second correlative words, both including nouns, verbs, adjectives and sometimes adverbs. For example, under the concept degrees of power there is first a set of analogous words whose meanings affirm the subconcept of strength (‘vigour’, ‘to force’, ‘strong’, ‘strongly’), and a second set (the correlative words) whose members negate the subconcept, e.g. weakness, fail, impotent. In the original layout of the *Thesaurus*, the two sets were presented in parallel columns, so that readers could find the word they were seeking under a positive or negatively related concept. Roget thus anticipated one of the findings of cognitive linguistics, that linguistic opposites are closely associated in the human brain (Aitchison 1992: 74–7). Roget’s parallel layout has been lost in modern versions of the *Thesaurus*, which simply interleave the two sets in successive paragraphs down a single-column page. Likewise Roget’s original designations ‘analogous words’, ‘correlative words’ are now overlooked,10 so that words grouped together in his *Thesaurus* sets are sometimes thought of as synonyms, though they may be synonymous in only one of their denotations and divergent in their connotations. There are of course no definitions to detail their meanings, as in a dictionary. Within the *Thesaurus*, word meanings are simply indicated by their place in the hierarchy of concepts, and their positive/negative polarity. This was in keeping with Roget’s aim of providing partial synonyms from which literate users might choose the most effective alternative for their purposes.

The original Roget’s *Thesaurus* included about 15,000 words, a relatively small inventory by comparison with modern dictionaries of the same size. It was however ‘greatly enlarged
and improved’ after Peter Roget’s death (1869) in two-yearly editions by his son John, and annual editions by grandson Samuel from 1925 to 1953. From c.1880 there were separate British and American editions, in which the British reviewed and enlarged the inventory within Roget’s original subcategories, and the American reworked the material and the overall hierarchy of concepts (Burger 1991). In Britain ‘Roget’s Thesaurus’ was trademarked, whereas in the US the Roget name became generic, and used by imitators and publishers of more and less thesaurus-like publications.

12.3.3 Ontological issues in thesauruses, especially in print form

Roget’s classification of English words by association with six superconcepts was designed to be exhaustive, so that each word could be found in its semantic niche within the hierarchical structure. Like other thesaurus-developers, he created his own multilayered semantic hierarchy (i.e. ontology) in the absence of any conventional model. The best-known examples of European thesauruses differ from Roget and from each other at the highest level: Hallig and Wartburg’s (1963) hierarchy for French has three supercategories: L’Univers, L’Homme, L’Homme et L’Univers, whereas Cesares’s for Spanish has just two: Dios, El Universo. However El Universo has many subcategories, for the organic and inorganic world, plants, animals, humans, individual mind/action/communication, society and its institutions (Ilson 2010: 248), making the structure very asymmetrical. These divergences in the conceptual hierarchies proposed for different European languages suggest linguistic and cultural relativities in the construction of thesauruses, which were overlooked by the universal language philosophers of the seventeenth century. The ontological construction of thesauruses for English also diverge considerably, according to whether they are designed for top-down access to the language at large, or developed ‘bottom up’ from a closed corpus or dictionary macrostructure (like the Historical Thesaurus of the Oxford English Dictionary, which generates a very large miscellany of superconcepts and over 200,000 subcategories). Either way the arbitrariness of the conceptual structure of the thesaurus is a disadvantage, making words less easy to find than Roget intended. In fact John Roget added an alphabetical index at the back of the family Thesaurus in 1879 (ten years after his father’s death), and it is now a standard feature with onomasiological thesauruses.

Roget’s Thesaurus and those modelled on it are notionally synchronic representations of the English lexicon (Ilson 2010: 251). Yet readers at the turn of the twenty-first century would be aware that some of the words contained in Roget’s sets of alternatives are rare or obsolescent, e.g. adumbration, alack, lachrymose, sedulity, tergiversation. This problem is addressed in the Historical Thesaurus of the Oxford English Dictionary (HTOED) published in 2009, which gives the first recorded date of each near-synonym and notes those which are obsolete. Drawing on the microstructural resources of the OED, HTOED is also able to provide stylistic information about the usage of words, e.g. booze as a colloquial term for ‘wine’, and plonk as an Australian/British word for ‘cheap wine’, though both are listed in a standard thesaurus without such labels, or definitions to help discriminate them. The onomasiological construction also masks the natural polysemy of common words, since the different senses of the same word have to be included in different subcategories, e.g. horse in both ‘animal’ and ‘framework’. There is no cross-referencing, so again it is the index rather than the ontological structure which provides this information.

Both synchronic and diachronic thesauruses are paradigmatic rather than syntagmatic (Ilson 2010: 256) in their projection of the lexicon. Neither shows the collocational tendencies that contribute to the meanings of words and help to discriminate their senses,
e.g. the different senses of the word freedom to be found in freedom from and freedom to. Other scholars have noted that thesaurus do not indicate the transitivity relations that verbs enter into, which help to differentiate transitive and intransitive/mediopassive uses (Burger 1991: 49–52). Compare ‘he sells books’ with ‘the book sells well’. The relative frequencies of words, now easily extracted from reference corpora and reported in learners dictionaries (e.g. Collins Cobuild and LDOCE), are not indicated in either the synchronic or diachronic thesaurus. While modern print dictionaries accommodate all these aspects of lexical meaning within the microstructure of individual entries, it is still unclear how/where they could be incorporated into the macrostructure of the print thesaurus.

12.3.4 Computerized lexical databases and the distributional thesaurus

Computer systems are game-changing for the thesaurus, helping to address many of the problems mentioned in the previous section. Their indexing capabilities allow instant access to elements at any structural level within the thesaurus, bypassing the non-intuitive conceptual classifications of the conventional print thesaurus. The pioneering online lexical database was WordNet, developed by psycholinguists at Princeton University, and grounded in research on the organization of the human mental lexicon (Miller et al. 1990). As in lexical memory, nouns are arranged in ‘topical hierarchies’ (p. 237), and accessed at more and less abstract levels in discourse production. This is matched in their layout, as shown below for bank.

(n) bank (sloping land (especially the slope beside a body of water))
(n) slope, incline, side (an elevated geological formation)
(n) geological formation, formation ((geology), the geological features of the earth)
(n) object, physical object (a tangible and visible entity; an entity that can cast a shadow)
(n) physical matter (an entity that has physical existence)
(n) entity (that which is known or inferred to have its own distinct existence (living or non-living))

WordNet thus constructs conceptual ontologies for individual words with rather more intermediate levels than the conventional thesaurus. WordNet was also designed to embrace all the senses of polysemous words (unlike the print thesaurus). It makes use of simple definitions and synonyms to identify the relevant lexical concept (Miller et al. 1990: 240), and each sense is illustrated in at least one corpus example, as shown below. The clusters of synonyms given for each sense of the word are ‘synsets’ – now 117,000 of them in WordNet version 3.1. The relative frequency of each sense is noted in brackets at the start of each line, based on its occurrences in the reference corpus:

(25) S: (n) bank (sloping land (especially the slope beside a body of water)) ‘they pulled the canoe up on the bank’; ‘he sat on the bank of the river and watched the currents’
(20) S: (n) depository financial institution, bank, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) ‘he cashed a check at the bank’; ‘that bank holds the mortgage on my home’
The features shown in those synsets are all very familiar from dictionaries, but lacking in Roget's Thesaurus. Yet WordNet is still essentially paradigmatic in its treatment of words and their senses, apart from what human users may glean from the examples.

The need for syntagmatic information on the behaviour of words is increasingly being met by the so-called distributional thesaurus, using data automatically extracted from large corpora (e.g. the internet) with targeted software. Distributional thesauruses compute the semantic similarities between common words, comparing the linguistic contexts in which they occur, and the parities in their collocations and grammatical dependencies. Distributional thesauruses have been developed for English and other languages, of which Sketch Engine is a widely known commercial example (Kilgariff et al. 2004), designed to serve the needs of artificial intelligence/knowledge abstraction in natural language processing. The ‘word sketches’ they produce of the collocational and grammatical behaviour of words are an asset in lexicographical research, and increasingly used as an adjunct facility in online general language dictionaries (Rundell 2012). The distributional thesaurus also serves the needs of translators and bilingual lexicographers, since parallel word sketches can be prepared for a word and its putative translation equivalent to see how closely they match up. They provide raw material for online bilingual thesauruses which could scarcely be realized in print.

12.3.5 Structured approaches to meaning in specialized lexicography online

Hierarchical structures of the concepts represented by everyday words (as in the WordNet illustration in §12.3.4) are also increasingly applied in specialized lexicography/terminography, and in information science. Sometimes called lightweight/lite ontologies (Giunchiglia and Zaihrayeu 2007), they are readily constructed using the hyponymic nomenclature that forms the backbone of the taxonomic sciences. Hierarchical relations among technical terms and concepts also serve the needs of artificial intelligence because they articulate ‘inheritance’ properties that can be exploited at multiple levels in the ontology. Yet twenty-first century descriptive terminologists (e.g. Temmerman 2000) recognize that technical terms are often polysemous and participate in more than one ontology, so that the properties they inherit will depend on the context. Technical terms also participate in non-hierarchical sense relations, e.g. meronymy (part–whole relationship), and in the relationship of instantiation, which can be represented schematically to explain disciplinary concepts in online terminography (Winston et al. 1987). Causal relationships and agency are crucial in the terminology of physical sciences and engineering (Faber et al. 2007). In architecture various types of associative meaning contribute to the multidimensional semantics of common terms such as window, e.g. their shape, location, function, constituent materials, all of which could be exploited in diagrams and tables, and in the underlying information structure of an online thesaurus (Fernández and Faber 2011). All these sense relations lend systematicity (as well as complexity) to the description of subsets of disciplinary concepts, but they can be displayed graphically by means of diagrams and drawings etc. to complement the verbal substance of the conventional dictionary/thesaurus (Peters et al. 2014). Specialized lexicography is naturally encyclopaedic (see §12.2.4), so the online specialized thesaurus that represents terminology in structured ways helps not only to profile the key disciplinary concepts but to model the construction of knowledge
in the discipline (Debnath et al. 2000). Similarly, the term *thesaurus* is now used to refer to the structure of information systems, like those used in libraries as well as the management of a bureaucracy.

### 12.4 Integration of dictionary and thesaurus

The distinctive construction of dictionaries and thesauruses – embodying semasiological and onomasiological approaches to the lexicon – have evolved separately over centuries. This chapter has traced the continual expansion of the macro- and microstructure of English dictionaries since the sixteenth century, contrasting with the late emergence of Roget’s *Thesaurus* in the mid-nineteenth century and its unique, fully articulated conceptual structure. Yet each has borrowed constructional features from the other. Learners dictionaries borrow from the semantic/conceptual approach, as does LDOCE (1995), in its cross-referencing from individual entries to summary pages or display boxes that present words with closely related senses: near synonyms, hyponyms or meronyms. Thus dictionaries can mitigate the tyranny of the alphabet when necessary. Conversely, thesauruses (from 1879 on) found it advisable to add an alphabetic index at the back of the book, to help users access lexical information within the not-too-intuitive conceptual structure. The two formats can usefully complement each other.

Fuller integration of the dictionary and thesaurus can be found in later twentieth century products such as the Reader’s Digest *Reverse Dictionary* (1989), where semantically related words (e.g. *gastronomy*, *cuisine*, *culinary*) are grouped under an alphabetically listed key word (*cooking*), all with brief definitions. The whole entry is cross-referenced to tables of more specialized words: *cooking terms*, *cooking utensils* and *menu terms*, to provide dictionary-like coverage of both common and ‘hard’ words. Major English dictionary publishers have used their lexical databases to create a dictionary–thesaurus in one volume, from the *Collins English Dictionary and Thesaurus* (1993) on. Some foreground the alphabetic index with brief definitions at the front of the book ahead of the thesaurus. Others e.g. *American Heritage Dictionary and Thesaurus* (2005) have the dictionary and thesaurus sections juxtaposed on double-page spreads, with the alphabetically ordered words and definitions on the upper half, and thesaurus categories in the lower half. The constraints of the print medium are evident in each case. Ongoing integration of the OED with HTOED in 2014 has the great advantage of working online from the start, able to link straight from the screen-ready alphabetical list with its detailed microstructure to the thesaurus entries, and vice versa. The online medium offers unlimited space for things that lexicographers could only dream of ten years ago (Schryver 2003), and for experimental combinations of verbal, graphic and audio elements that can be tested to optimize their use.

Innovative combinations of dictionary and thesaurus online are still governed by established lexicographical principles (Rundell 2012). Planning the macrostructure still matters, in terms of what parts of the lexicon – words and multiword units, common and specialized words, proper nouns and idioms – are to be included. There are new constructional opportunities in making use of the mediostructure, not only to cross-reference from word to word, or from dictionary to thesaurus, but to connect specialized terms with encyclopaedic information elsewhere on the internet. Links to selected corpus data and the distributional thesaurus can enhance the collocational information offered in online learners dictionaries. More grammatical information can be offered than could be squeezed into the print dictionary. But with increased internal and external linking, navigation and systems for content management become ever more important. The allied disciplines of information
technology and artificial intelligence are there to support computer-searches of words by form as well as their semantics, for both human and machine readers. The computerized microstructural template is more capacious than in ordinary print dictionaries, but still raises the question of what to prioritize on the first screenful, displayed landscape-wise rather than portrait-wise.

These twenty-first century aspects of lexicographic practice, and questions of how to optimize the functionality of the combined dictionary–thesaurus for human users, require continuing research to enlarge the empirical foundations of lexicography. The principles that have evolved in dictionary-making and construction now need to embrace both semasiological and onomasiological approaches to the lexicon, to support any fully-fledged lexicographical theory.

Notes
1 The plural form of thesaurus is either neo-Latin thesauri or English thesauruses. In the original Greek, the plural was thesauroi. Since using thesaurus to refer to a semantically structured model of the lexicon dates from the nineteenth century (see §12.3.1), it seems reasonable to use the English plural in this chapter’s title and elsewhere.
2 Cawdrey’s title uses the word ‘usuall’ in its earlier sense of ‘current’ (see Stein 2010).
3 The intent of Wilkins’s work is obscured for twenty-first century readers by its use of ‘essay’ (=experiment), and ‘real character’ to refer to a symbolic system that stands for universal concepts (see Claus 1982).
4 These terms originated with Darwin (1857) in biological taxonomy, according to the Oxford English Dictionary online.
5 The ‘hard words’ tradition is still exemplified by Ayto’s Dictionary of Difficult Words (1993), published by Helicon.
6 The initials JK probably refer to John Kersey, the author of other early eighteenth-century dictionaries (Starnes and Noyes 1946: 69–70).
7 The OED included adjectives (and verbs) derived from proper nouns, e.g. Australian, but not proper nouns like Australia, a policy which presented challenges (Murray: General Introduction). By not including proper nouns, the OED was spared the problem for encyclopaedic dictionaries of continually updating facts contained within the microstructure.
8 IPA is the International Phonetic Alphabet for rendering the individual sounds/phonemes of any language, independent of its orthography.
9 The LawTermFinder online termbank in Australian family law includes external links to the relevant legislation.
10 Except in Merriam–Webster Dictionary of Synonyms (1984), where they are systematically used.

References


