Dictionary and thesaurus in the digital environment

David Crystal

Integration of a dictionary and thesaurus was a physical impossibility in a paper world, except on a very small scale, as in a thematic glossary. In a digital environment, the interaction between OED and HTOED is practicable, dynamic, and fruitful. But, as an unprecedented exercise, it is very much a first approximation to the ideal, and one of the things we need to do is identify the ways in which user experience can be improved.

For a forthcoming OUP book (*Words in Time and Place: Exploring the Historical Thesaurus of the Oxford English Dictionary*, 2014), I examined 15 semantic fields within the HTOED taxonomy. The aim of the book is to introduce the general reader to the wealth of data in the thesaurus, but during the course of the work a number of issues arose which need to be addressed in long-term planning, most of which are to do with the interaction between the notions of taxonomy, semantic field, and etymology.

The notion of taxonomy

Taxonomies are traditionally encyclopedic, not linguistic, and thus go well beyond the notion of semantics as linguists use the term. The ultimate online encylopedic experience is intended to be the Semantic Web (SW), but we should note that *semantic* here no longer means 'linguistic meaning' but 'all knowledge'. For example, the SW will (in its anticipated full development) tell you not just the difference in meaning between a *hotel* and an *inn* (two HTOED categories), but where the hotels and inns are on your next holiday trip, how to get there, whether there is availability, how much rooms cost, and so on - information that does not interest lexicologists at all (unless they are on holiday).

However, semantic taxonomies (in the linguist's sense, such as Roget and HTOED) rely totally on an encyclopedic perspective. All provide a 'top-down' classification based on knowledge, not semantic relationships. HTOED begins with three very general categories (The External World, The Mental World, The Social World), and makes increasingly specific subdivisions until editors decide no further distinctions are worth making. These categories and their subcategories are nonlinguistic in character, based on real world criteria (such as biological species, historical periods, and geographical divisions), applicable to all languages, and - of especial importance - reflecting the mindset of the designer. There is no such thing as a universally recognized 'logical' taxonomy, and any exploration of the HTOED classification needs to bear this in mind.

The best way to show that taxonomic mindsets exist is to compare them. For example, the taxonomy found in the Dewey decimal classification system, widely used in libraries, differs in many ways from that used in other systems and in the HTOED. Dewey's 'top ten' categories (general works, philosophy, religion, social sciences, language, pure science, technology, the arts, literature, history) very much reflect the interests and beliefs of its author. Compare these with the top-ten categories in the Global Data Model I devised for internet search, now used widely in online advertising (under the brand-name of iSense): the universe, the earth, the environment, natural history, humanity, recreation, society, the mind, human history, human geography. And as one looks at lower-level categories, differences multiply. To take just one example: in Dewey, Central America is listed as part of North America; in *HTOED* it is grouped along with South America. Thousands of quite basic questions have to be decided, and taxonomists make different decisions (is Greenland a continent? is Turkey in Europe? is a tomato a vegetable? is photography an art? what counts as a craft? is a satellite a spacecraft?). There is inevitably a certain amount of arbitrariness and personal taste in decision-making, and users of any thesaurus need to be aware of it, as it affects where they look and what they will find (or not find) there.

Missing item syndrome

The biggest problem facing anyone using a taxonomy of lexical items is what I would call 'missing item syndrome'. We go to a particular semantic category and see a listing of synonyms - only we notice that some items we expect to be there are missing. We look in the index and discover that the item is definitely in the thesaurus, but in an associated category - either 'vertically' (higher up or lower down the classification, as with relationships of hypernymy and hyponymy) or 'horizontally' (at the same level of the classification, as with antonymy and incompatibility). This problem affects HTOED as any other thesaurus, but in HTOED it is present on an unprecedented scale, given the large number of lexical items involved and the historical dimension of the work. I am not sure how to solve this, other

than by applying a structural semantic model in a stronger way, and taking etymology more fully into account. Let me illustrate what I mean with two examples from HTOED.

The problem of vertical categories

In a taxonomy, we normally expect the range of the semantic categories to reduce as we go further down the hierarchy. Beginning with X, we expect the next level down (X1) to be aspects of X, and the next level down again (X2) to be aspects of X1, and so on. But sometimes there is overlap between higher and lower categories, as can be illustrated by the HTOED listing for 'privy' (see Wordmap 1 below). The basic idea behind the classification at this point is that the subcategories deal with special cases. Thus within the general category of 'privy/latrine' we see such special settings as 'in a ship', 'in a convent', and 'used by a king', or special pieces of equipment such as 'urinals', 'chamber-pots', and 'bed-pans'. The subcategory 'water-closet/lavatory' focuses on the room in which the lavatorial facility is situated, rather than the facility itself. Sometimes this distinction is clear-cut. There is no ambiguity with such words as *restroom* or *bathroom*. One always 'goes to' these places, and never 'sits on' them. But not all the words are like these.

For example, the *OED* definition of *water-closet* begins 'a closet or small room fitted up to serve as a privy'. This seems clear enough until we read the usage note: 'sometimes applied to the pan and the connected apparatus for flushing and discharge; also, loosely, to any kind of privy'. For semanticists and taxonomists, the devil always lies in such details - here, in those words 'sometimes' and 'loosely'. The multiple applications of words like *water-closet*, made more complex always by the vagaries of language change and uncertainties over usage, both socially and regionally, means that there are bound to be arbitrary decisions over which category to assign a word to (as with *little house*, which has been put in the category 'privy/latrine', and *smallest room*, which has been put in 'water-closet/lavatory'). If the taxonomist feels that the 'bowl + pedestal' (or equivalent) is the dominant element, the word will be assigned to the first category; if 'the room in which this is situated' is felt to be the dominant element, it will be assigned to the second. In several instances, everything depends on the context, and even on the choice of grammatical construction, as we see with such entries as *lavatory*, *toilet*, *loo*, and *john*, which allow both 'going to' and 'sitting on'.

The problem of horizontal categories

This can be illustrated from the semantic field of 'stupid/foolish/inadequate person' (See Wordmap 2 below). Clearly a potential distinction can be drawn between the two extremes of 'blockhead' (someone with their senses intact who is acting stupidly) and 'simpleton' (someone with a weak intellect). 'Fool' hovers somewhat uncertainly in between. And when we examine actual contexts of use for the various words, it is often difficult to say which of these three emphases is dominant in an individual citation. As a result, some sideways cross-referencing between the categories is inevitable.

If users restrict themselves to only the words in the 'fool' list, there would be several cases where the etymological story would be only partly told. For example, look up *nigion* (1570) in the *OED*. The entry states '= NIDGET *n*.', and the cross-reference shows that this is indeed a relevant part of the explanation of the word. But *nidget* is not in the 'fool' listing: it is in one of the divisions of the 'simpleton' category. Similarly, the etymology of *silly ass* (1901) needs a reference back to both *silly* and *ass*: *silly* is in the 'fool' list, but the etymological reference to 'ASS *n*. 2' (1578) takes us to the 'blockhead' category. *Goof* (see also *goff*), *niddipol* (see *noddypoll*), and *gump* (see also *numps*) are further examples of the need to go sideways through the thesaurus categories to complete a lexical history.

In addition to the 'horizontal' examples just illustrated, there are several 'vertical' missing items in the 'fool' category. *Gobdaw* (1966) obviously needs to be linked to *daw* (c1500), but that is in a *subcategory* of 'fool'. And similarly in this subcategory we will find *goose* (needed to complete the story *of saddle-goose* and *goosey*), *noodle* (needed for *doodle*), *noddy* (needed for *nodcock*), *hoddypeak* (needed for *noddypeak*), and *ninny* (needed for *nincompoop*).

The scale of the problem needs to be appreciated. There are over 100 items in the semantic field of 'fool', but these need to be seen alongside the additional 40 in the 'weak intellect' category (*sucker*, *dope*, *softy* ...), the over 200 in the 'blockhead' category (*nitwit*, *moron*, *thicko* ...), and the 70 in the further subcategory of 'fool'. In all we are dealing with the classification of over 400 words of an extremely colloquial kind. There are hardly any formal words in the list (*insipient*, *foolane*, and *liripipe* are exceptions): people are not usually being formal when they refer to each other as fools. But the more colloquial the words in a semantic field, the more difficult it is to categorize them uniquely.

Theoretical questions aside (eg whether a lexical item stays within the same semantic field throughout its history), what these two illustrations mean is that users of a thesaurus must always be

prepared to look upwards, downwards, and sideways when exploring a category, especially one where subcategories are closely related. They need to read (or at least, skim) through the whole of a semantic field before deciding to focus on a part of it. At the moment, this is not easy to do. In the paper edition, one keeps having to move between Volume 2 and Volume 1. Online, it is not easy to view different thesaurus categories simultaneously. There is room for presentational improvement here.

Lexical comprehensiveness

Traditional taxonomies have never been lexically comprehensive, partly because of the limitations of paper publishing, and partly because, when dealing with everything, it is impossible to keep the material up to date. We might think that online taxonomies would have solved these problems, but we would be wrong.

Synchronic coverage
Online taxonomies are by no means lexically comprehensive. They are usually designed for a particular and limited purpose, such as a classification of what is available in a retail store, and they have not been designed by people who have much lexical awareness. During the iSense research we repeatedly encountered remarkable omissions. I once went to a major online retail store, typed in *mobile phones* into a search box and got the response 'we have no mobile phones'. I tried all kinds of variants (*cellphone*, *cell-phone*, *mobile*, *mobile telephone*...) until I found the one the system recognized (in this case, *cellular phones*). Online search has become a bit more lexically sophisticated in the last ten years, but it is still a long way from the kind of comprehensiveness a good dictionary would provide, such as allowing for variations in formality, capitalization, hyphenation, and regional spelling. This is where an OED lexicology can be hugely informative, as in principle it includes all spelling and stylistic variants. And, looking ahead, pronunciation variants too [see newsletter item, reproduced below].

Diachronic coverage
Language change is not just a phenomenon of the past; it is ongoing, and its effects need to be monitored. Traditionally, lexicographers do this by relying on their daily listening and reading (or that of their contributors) to notice lexical change, and this procedure is generally sufficient to capture most of what is going on. But a thesaurus changes the goalposts, as the encyclopedic dimension is critical, and this introduces dimensions of change that British dictionaries have generally avoided, and that American dictionaries have handled in a very limited way (including only important people and places). I am talking here about all the proper names that a comprehensive thesaurus would include, and it is here that this genre is most challenged, as it has to keep pace with the flow of new products (and new versions of products) coming out of industry, the yearly (or subyearly) changes taking place in the members of sports teams, the changes in political leadership in countries, and a huge number of expressions derived from these names (such as 'It was like Clapham Junction in there', 'My watch is more Portobello Road than Bond Street', 'Rolls Royce or Reliant Robin?' [a newspaper headline, but not about cars], 'Ben is the Jamie Oliver of Shakespeare', and 'Disgusted of Tunbridge Wells'). (This point overlaps with the issues that arise in relation to coverage of global English, as few of these expressions 'travel', and different cultures have their own versions.)

How does one monitor all this? There is so much of it that we need to look for an automatic solution, such as some sort of 'data pump'. This term has several interpretations, but here I mean a system which monitors graphic units (or strings of such units) on the internet and 'pumps out' - on a daily, weekly, or whatever basis - lists of items not previously included in the database, and sends these for consideration by a human editor. A level of granularity may have to be decided by the editor, so that items which come above that level (eg a new type of camera or a new genre of pop music) are included and those which fall below that level (eg a new model of an old camera or a revival of an old musical genre) are disregarded. You might think this is an impossibly large task; in fact, it turns out to be relatively straightforward, as only a few areas of knowledge change lexically[1] very rapidly, and an experienced taxonomist knows what to look out for.

[1]*By lexical change, I mean novel items or combinations of items. I am not talking about the whole of human knowledge of a Mastermind type - 'Who won the FA Cup in such-a-year?' 'What is the largest lake in Cumbria?' - but only of the lexical items that are needed to allow such questions to be asked and answered. A Thesaurus is not a Ready Reference.*

Integration and citation

The incorporation of HTOED into the OED raises an issue which no lexicographer likes to think about: the need to check the accuracy of earlier editors. The book I mentioned earlier contains 1243 entries that tell the history (in a very minimal form) of the items listed in the 15 selected categories, illustrated by *OED* definitions and citations. The good news? I checked every citation against the original text (the vast majority are now available online, making this task infinitely easier than it would have been a few years ago), and found only one case where the author was wrong and one case where there was a wrong line reference. The bad news? There were 33 cases where there was an error in the dating cross-reference. This is 2.7 percent, and that is an uncomfortably high figure, and clearly some sort of checking strategy needs to be introduced as revisions proceed.

What kind of errors are they?

- most involve cases where the HTOED entry gives a first citation date that is wrong, because it occurs as part of a list of related expressions, and the software has selected the first citation in the list, which is not the correct one: for example, in the category *prostitute*, *soiled dove* shows a link to OED *soiled* (adj.1), and gives the first citation as 1250; but the 1250 citation is to 'soiled lips' and the relevant 'dove' citation is actually 1882.
- the cross-reference is to the wrong sense of a word: in the category 'drunk', *rummy* 1742 takes us to the first citation in sense 1 of *rummy* (adj.1), whereas it should be to the first citation (1834) in sense 2.
- the cross-reference is to the wrong entry: in the category 'drunk' (adj.) we find *topped* 1632, which is linked to *topped* (adj.1), but neither of the senses listed there are anything to do with drinking; rather, we have to go to *top* (v.3, sense 2) to find the relevant 1632 citation.
- in multi-word expressions, there may be a discrepancy between the first recorded use in the different words: in the category 'die' (v. intrans), *pop off the hooks* is linked to *pop* 1764; this should in any case be 1887 at *pop off*, but if we go to *hook* (n.15e) we find an earlier usage, 1842.
- the dates get associated with the wrong item: in the category 'die' (v. intrans), *slip one's wind* is said to be 1819, and the link sends us to *wind* (n.1, 11.c) where the correct citation is actually 860, and the 1819 citation there is irrelevant; only if we go to *slip one's breath* (at *slip* v.1, 26.c) do we find the relevant 1819 citation
- an entry is in twice, in the superordinate field and in the subordinate field: in 'popular music', there is a subcategory *heavy metal* in which there is a link to the OED entry *metal* (1984); but if we go to the subcategory of *heavy metal*, *types of*, we see the same link to *metal* (1984).
- an entry is in twice, in two separate subcategories: in 'popular music', in the subcategory of *rock*, *types of*, we find a link to *folk-rock* (1966); but if we go to the subcategory *other pop music*, we find it there too (though with the wrong cross-reference, to an irrelevant 1965 citation).
- a link leads to a completely irrelevant entry: in 'prostitute', there is a word *pliers* 1678, which takes us to *pliers* (n.), but none of the ten citations at *pliers* have any such figurative illustration; it takes a feat of imagination to use the 1678 quotation in this way, which is a quotation from a book by the globe-maker James Moxon called *Mechanick exercises: or, the doctrine of handy-works*, who explains: 'Plyers are of two sorts, Flat Nos'd and Round Nos'd. Their office is to hold and fasten upon all small work, and to fit it in its place.'

Wicked sense of humour, these lexicographers.

Appendix [Published before the symposium in the OED Symposium online Newsletter]

We ain't seen nothin' yet

David Crystal

I once ended a book on internet language by saying 'we ain't seen nothin' yet'. I was right, for I wrote that before Twitter arrived! After a mere couple of decades, it has to be acknowledged that the internet is still in its infancy, and we must expect radical changes, some of which will affect the OED. Here are two.

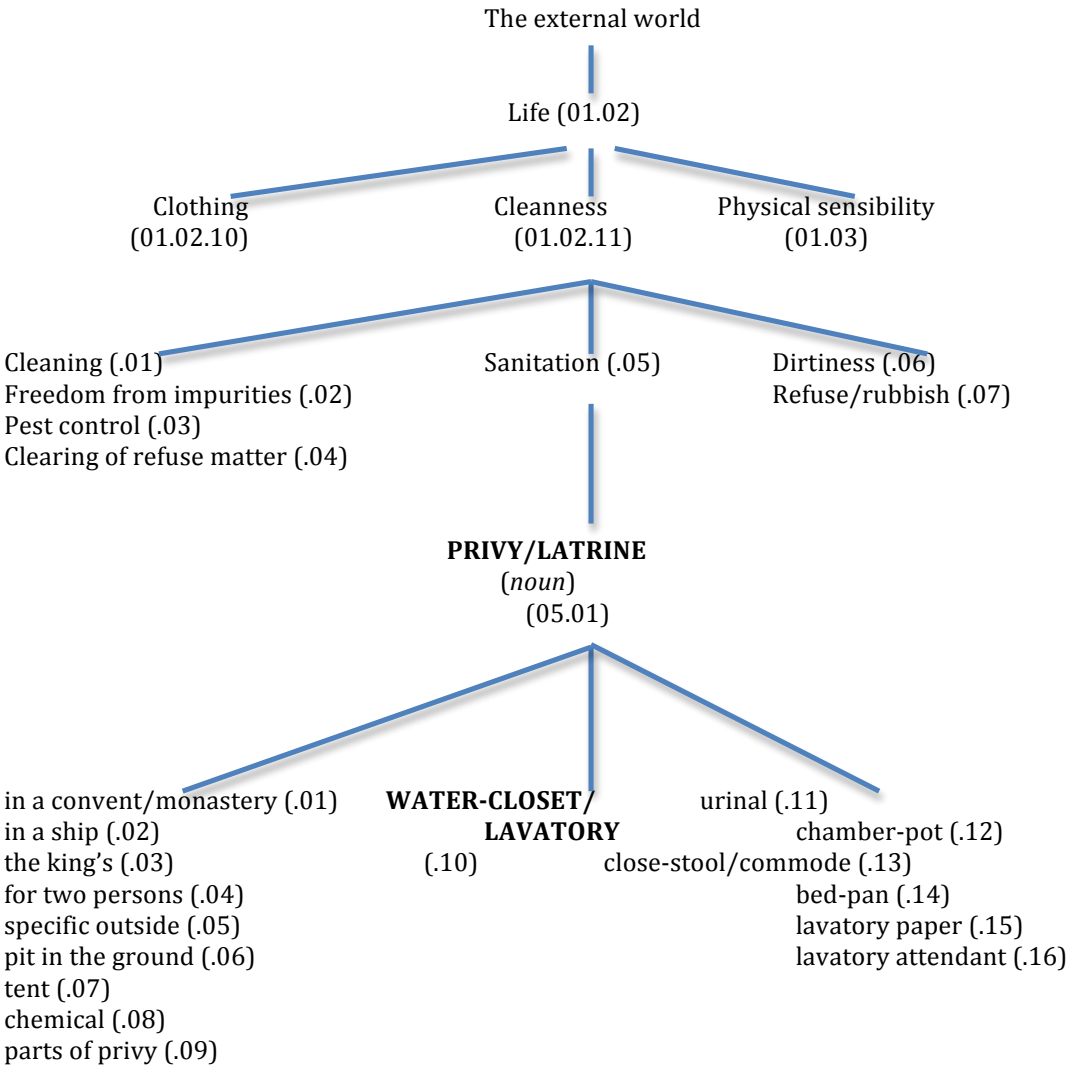
Miniaturization. At a mobile technology conference I attended recently, it was being asserted that by 2020 some 80 percent of internet access is going to be by mobile phone. It already at this level (or higher) in parts of the world where wired technology is unavailable (as in many parts of Africa). But what happens to information when it is compressed into a small screen? I don't see much of a problem with an enquiry about individual OED lexical items, but a thesaurus classification requires a large screen to see the relationships between sets of categories, and the lexical lists which emerge within a particular category are sometimes very large (over 100 items). How will these be best presented on a small screen? A great deal of attention has already been paid by the OED team into the best way of presenting lexical information on screen, and this will need to continue. Testing accessibility and legibility in a mobile environment is perhaps not a priority at present, but it will be one day.

Audiovisualization
The internet is currently a largely graphic medium; only about 20 percent is audio or audio/visual. But this is about to change. Speech-to-text and text-to-speech technologies, and multimedia streaming, will reverse this proportion in the next decade or so. Already, with many systems (such as GPS and automated phone enquiries) audio presentation is the norm. There are two developments which need to be anticipated.

Speech to text: we must anticipate a day when we do not type an item into an OED search box, but speak it in. This will avoid the problem of people mistyping or not knowing how to spell a word. At present, speech to text systems have several limitations which reduce their value (difficulties with regional accents, fast speech, interference from background noise, and proper names), but these will become less as time goes by. For a dictionary, the question of how to deal with homophones needs to be anticipated, as well as how to handle words with similar pronunciations ('sounds like this').

Text to speech: this is far ahead of speech to text in terms of practical applications. Several dictionaries and online sites already offer audio pronunciations, rather than relying solely on phonetic transcription. This has to be the way forward with the OED - but with a similar need for comprehensiveness, with respect to major international variants (eg British vs American), gender balance (male and female), stylistic variants (especially formal vs informal, but also sometimes occupational), and - to my mind most interesting of all - historical (the 'original pronunciation' movement, where people want to hear speech as it sounded in its period, insofar as this can be established, which is now attracting a lot of interest worldwide). Early pronunciation is covered very patchily by the OED, and reflects the philological orientation of the early editors rather than the kind of thing we find these days in historical phonology.

Wordmap 1 (adapted from *Words in Time and Place*)



Wordmap 2 (adapted from *Words in Time and Place*)