How to keep up with language change? Have your head in the clouds.

How can dictionaries cope, faced with rapid global and technological change? Partly by improving coverage - adding new words; partly by improving treatment - handling words better. Words are not simply individual items to be learned; they form 'clouds' of structured meaning. Notions such as collocation and thesaurus become central. Illustrations from the new edition of the Longman Dictionary of Contemporary English.

'Unhappy mortals', said Dr Johnson in the Preface to his *Dictionary*. He was talking about those people who, as he put it, 'toil at the lower employments of life ... where success would have been without applause, and diligence without reward'. You might think he was referring to language teachers, or even linguists, but he was not. He was talking about the writers of dictionaries. He tells us why they are so unhappy a little later in his Preface. Language is so volatile, he says, that trying to constrain it, as academies hope to do, is futile. It is like trying 'to lash the wind'. Or, more appropriately for the present talk, as we shall see, the clouds.

Language change is the problem. How on earth is a lexicographer to keep up with it, especially in an era when English is changing more rapidly than at any time in its history. Two forces are driving this change. The first is the global spread of English, which has added at least 100,000 words to the language in the past few decades, judging by the totals included in recently published regional dictionaries. Not that all these words are neologisms, of course. For the most part it is simply the first time they have been recorded in a dictionary. But, having been recorded, they are now publicly accessible, and it is accessibility that drives change. We cannot adopt a new word or meaning until we know it exists.

And how do we know it exists? This is where the second force comes in: the power unleashed by the internet. It is now possible to access the local lexicons of the English-speaking world more easily than ever before. Once upon a time, if I wanted to find out about South African English, as represented, say, in its daily newspapers, I would have had to go to South Africa. Now I call them up online. I can encounter informal written South African English in innumerable forums, chatrooms, blogs, and social networking sites. The same applies to other parts of the world. This immediately increases my passive vocabulary; and in some cases it will increase my

active vocabulary too. If the notions involved are of interest outside of the originating country, the words will creep into standard English. *Sudoku*, added in *LDOCE5*, is a case in point.

In fact, *creep* is the wrong word. New words and expressions encountered on the internet are capable of entering our mental lexicon more rapidly than at any time in the past. If I invent a new word today, and put it in my blog, it can be around the world in next to no time. There may even be evidence of lexical spread, in the form of the comments that people submit to the blog, which often show that they have picked up the new usage. People read my blog, according to the user logs, in over 100 countries. Comments can arrive almost as soon as the blog appears, as many users have a system which alerts them to new posts straight away. There has never been anything like this before, in the history of lexical transmission. So we definitely need a better descriptor than *creep*. I went to *LDOCE5*'s thesaurus feature at *run* (slide 1). Yes, words travel around the internet at a pace: they *race*, *dash*, *sprint*, *charge*, *tear* around.

The internet has so far had a limited role in relation to language change. It has added new genres, such as email and text-messaging; new styles - notably, at the informality end of the formality spectrum; and new orthography, illustrated by new functions of punctuation in e-addresses, and by emoticons. But the impact on lexicon and grammar has been minimal. If we collect all the new words and phrases which have entered English as a result of the internet - such as *mouse, click, blog, text*, and the like - we are talking hundreds, not thousands. It's not too difficult to keep pace with them. Several have been added to *LDOCE5*: of the words which were not included in *LDOCE4* you will notice *crackberry*, *video blogging*, *blogzine*, *webinar*, *webzine*, and *social networking site*, as well as the names of some of these sites (*Facebook*, *MySpace*, *YouTube*, *Second Life*) and other online enterprises, such as *eBay* and *Skype*. There are new senses, too, such as *piggyback* (to use someone else's connection to the Internet without them knowing), *poke* (to let someone on a social networking site know that you want to communicate with them), and the verb use of *friend* (make someone a friend on a social networking site). But if we add all these new items up, we're unlikely to reach more than about a thousand or so. And that is a drop in the ocean compared with the size of the English lexicon as a whole, which comprises well over a million lexical items. Lexicographers can handle this. This is not where unhappiness lies.

No, the problem comes not in relation to coverage but in relation to treatment. These are the two big dimensions in lexicography. Coverage attracts all the attention. When a new edition of a dictionary comes out, the press generally pick up on the new words it includes, as if this were the be-all and end-all of a dictionary. Certainly, coverage is an important part of the story. We only have to notice the new words included in *LDOCE5* to obtain an insight into the nature of the social forces motivating language change - *biodiesel, biodigester, biofuel, biohazard, biosecurity; carbon credit, carbon footprint, carbon neutral, carbon offsetting.* But we should never judge a dictionary by the number of new words it includes in a new edition, which are always going to be relatively few - in a dictionary of this size, rarely more than a hundred with each edition. Rather, judgement should focus on the other dimension, treatment.

Treatment means, quite simply, how the editors handle an entry. Simply look at an entry and you will see the treatment - the graphic design, the headword information (about grammar, pronunciation, variation), the senses, the examples, the cross-references, the special features. It would be possible to discuss lexicographical progress in relation to any of these topics. Today, I want to focus on the two new special features of *LDOCE5*, as they do I believe illustrate the direction in which lexicography is moving - or needs to move: the thesaurus panels and the collocation panels.

All dictionaries are trying to do the impossible. They are trying to capture the multidimensional nature of the lexical semantics we have in our heads by imposing a unidimensional discipline on the page that is little suited to the task - alphabetical order. It's a huge convenience, of course, but it totally destroys our sense of semantic structure. *Aunt* is at one end of the dictionary; *uncle* is at the other. Yet we all know that aunts and uncles should lie together. This one is easy to solve, of course, by using a cross-reference. All we need to so is have *uncle* > *aunt* (which *LDOCE5* does) and *aunt* > *uncle* (which, curiously, it doesn't). But most of the semantic relationships between words are not so simple. There is no neat binary opposition. Rather, the words form 'clouds' of structured meaning.

The concept of the *word-cloud* has become popular of late, thanks to developments such as Wordle. Wordle finds all the words in a text and presents them as a visual cluster, with the frequency of the words related to the amount of visual prominence they receive. They are a good way of gaining an immediate impression of

the content of a website or text. And they can throw up surprises. Process the textual content of *Othello*, for example, and who do you think will be the most referenced character? Othello? No. Iago? No. It is Cassio. It makes you think. Wordle is just one type of cloud presentation. *Data clouds* can be based on other things than frequency, such as stockmarket prices. A type of presentation that linguists find useful is the *collocational cloud*, which plots the words most often used in association with a particular target word.

These are excellent visual aids, and they provide informative impressions of the way language is being used. (See the Wordle of my lecture at slide 2.) The limitation is that they tell us nothing about meaning. This was the problem with the traditional thesaurus, of course. A thematic thesaurus, such as *Roget's Thesaurus*, or an alphabetical one, such as *The Cambridge Thesaurus of American English*, simply lists words in sets (slide 3). If you want to find the meaning of one of these words, you have to go to a dictionary. Another limitation is that the sets are linguistically unprincipled or totally unstructured - sometimes simply alphabetical lists, often random aggregates. The sets can also be extremely large - too large for practical use in learning situations. A thesaurus entry in *Roget*, for example, can consist of over 500 items.

There have, accordingly, been several attempts to combine the properties of dictionary and thesaurus. One of the first was the *Longman Lexicon*, which Tom McArthur edited in 1981. And in *LDOCE5* we see the latest attempt to do both. The thesaurus panels consist of small groups of semantically related items, such as the *WALK* set (slide 4). Each entry gives a definition which usually either replicates or is a paraphrase of the entry in its alphabetical place, and illustrates this with a corpus example (different from the one(s) used in the main entry).

Note that this can't be an automatic process. You can't just take the definitions and examples from the headword entries and jam them together into a thesaurus panel. The contrasts required by the definitions in a thesaurus compilation may need to be rephrased to enable users to see how the senses complement each other. They also need to be more succinct. It is essential to view the definitions and examples together, otherwise you might feel the dictionary is contradicting itself. *Wade*, for example, is defined in its alphabetical place as 'to walk through water that is not deep', but in the thesaurus panel as 'to walk through deep water'. Both senses are of course possible, depending on what it is that is being waded through - in the thesaurus

example, *We had to wade across the river*. No example is given in the main entry, but if one were supplied it would presumably be something like *I waded through the puddle in my wellies*. This is where the accompanying CD comes into its own, of course. Look up *wade* there and we see *wade into the surf* and others.

In structural semantics, the sets of words in thesaurus panels would be called *incompatible terms*. They are not strictly synonyms; rather, they provide semantically distinct alternatives. Under *WALK*, we can *wander* or *stride* or *pace* or *march* or *wade* or *stomp*, but we can't do two of these in the same action. If we're *marching*, we're not *wading*, and vice versa. There are of course other structural semantic relationships which we need to know about, such as various types of opposite, or the relationship between part and whole. In an ideal multi-dimensional dictionary, these would be present too. No dictionary, not even *LDOCE*, is able to do everything. Earlier on I referred to words *creeping* across the internet, and suggested we needed a faster notion. Somehow we have to get from *creep* to *run*. In *LDOCE5* the thesaurus cross-reference from *creep* takes us to the *WALK* panel, and we find creep there under *WALK SLOWLY*. But there is no explicit link taking us to the thesaurus panel at *RUN*. The inclusion of thesaurus antonyms is a development for the future, perhaps - though it remains to be seen whether it's practicable to include on an already graphically rich page yet another dimension of structural semantic information.

Another way of breaking through the barrier of alphabetical order is by focusing on collocations. Collocations are the mutual sequential predictabilities between words. What is the likelihood of X preceding Y or of Y following X? Note that the two questions are not the same. *Amok (amuck)* has to be preceded by *run*, but *run* does not have to be followed by *amok* - we can *run a mile, run water*, and run all kinds of things. It's very important to look in both directions, as can be seen in the *LDOCE5* collocations panel at *BUS*, where we see *bus ride/stop/shelter etc.* on the one hand, and *school/shuttle/double-decker etc. bus* on the other (slide 5). Using this technique, it's also possible to become aware of important changes of meaning. Take the panel on *CHILD* (slide 6). In the cases where it is used with a preceding adjective, the meaning can switch from singular to plural (*young/small/gifted etc child*(*ren*)); in the cases where it is itself used adjectivally (*child abuse/development/labour*), the meaning is always plural, and *children* is disallowed.

The problem for the lexicographer is that every word has collocations, and many words have an indefinitely large number of collocations, running from the ones

that immediately come to mind to those which are more occasional. It is, if you like, a scale of idiomaticness, ranging from word sequences which are totally idiomatic, allowing no change (as with *run amok* and *spick and span*), to those where some sequences are preferred over others. If we look at the set of words in the *LDOCE5* thesaurus panel at *BEGINNING*, we find *beginning, start, commencement, origin, onset, dawn*, and *birth* (slide 7). How far will these be used after, say, *auspicious*? Only the first two are likely, as a quick search in Google would show, but that search engine brings to light a sprinkling of hits for all the others too. The problem for the lexicographer is always deciding where to draw the line. In the present example, the answer is easy: only the collocations with *start* and *beginning* have high frequencies (around 80,000 hits in Google), and so it is not surprising to see that these are the two specifically mentioned at the entry on *auspicious*. The others have hits of around a thousand, or, in the case of *onset* (on the day I looked) a measly eight hits. We would not expect them to receive special mention in a learners' dictionary. In other cases, the line-drawing is much harder - for example, deciding how many collocations to allow in for words like *occasion* or *efficiency*. I suppose the principle here is that any information is better than none. Traditionally, dictionaries provided none. The collocation panels in *LDOCE5*, and the lists on its CD, point the way forward.

Lexicographical research into thesauruses and collocations evidently has three aspects. First, we have to establish the facts, using corpora. Second, we have to present the facts, using whatever graphic and electronic means are available. And third, we have to keep the account up to date. It is this last point which is the bugbear. Imagine, The unhappy mortals spend many lexicographical hours establishing a lexical system, and many more presenting it elegantly so that it fits perfectly on a page - and then the language changes and messes everything up. New words or senses have to elbow their way into already existing lexical sets. Old collocations disappear. New collocations emerge. No wonder they remain unhappy.

Old collocations disappear? We only have to look at the pages of the unabridged *Oxford English Dictionary* to see this. Take the verb *cast*, collocating today (as one goes through the list of senses in *LDOCE5*) with such words as *light, doubt, shadow, glance, eye, vote, spell , mind*, and *aspersions*. We no longer *cast a chance*, or *a ditch*, or *reckonings*, or *water*, or *love* - just a few of the collocations for this verb in past times.

A good example of an impending thesaurus change is the *LDOCE5* set at *TALK*, which contains *talk, speak, go on/drone on/ramble, waffle, prattle on*; *TALK ABOUT EVERYDAY THINGS have a conversation, chat, gossip, visit with, converse*; and *TALK SERIOUSLY discuss, talk over, debate* (slide 8). The arrival of the internet must soon alter this lexical balance, where a new category, such as *TALK ELECTRONICALLY*, will need to include *chat* (in a different sense), *blog, text, message* (as a verb), and others. And good examples of collocational change can be seen from the new words already included in this dictionary - individual collocations such as *hybrid + car, happy + slapping, helicopter + parents*, and *credit + crunch*, or collocations involving lexical sets, such as the economic terms collocating with *green* (*audit/tax*, etc), *emission* (*credit/trading*, etc), and *carbon* (*footprint/offset*, etc). The internet is here too, as we see with new collocations such as *bunch of text*. New collocations have no order and no end. And the shockwaves which follow their arrival sometimes extend into several areas of the lexicon.

So, it seems, lexicographers are likely to remain unhappy mortals, as they continue trying to lash the lexical wind. At the same time, I am sure they must get some satisfaction from the fact that the increasing availability of linguistically informed dictionaries, of which *LDOCE5* is an exemplar, is actually making a lot of other people very happy indeed.