Towards a superdictionary

This is the text of a (hitherto unpublished) paper I delivered as the inaugural Michael Samuels lecture at the University of Glasgow in October 2012. It was called 'From super dictionary to super-dictionary', and it gives more details of the issues involved in quantifying the English lexicon.

On 15 January 1965, at a meeting of the Philological Society in London, Michael Samuels made his first public announcement of the Historical Thesaurus project. It took the meeting somewhat by surprise. His paper was on 'The role of functional selection in the history of English', and was a standard philological presentation – a detailed discussion of the factors involved in language change and the need for an integration of the various models into, as he put it, 'a single, all-embracing theory of diachronic linguistics' (p. 17). I was at that meeting, and the last thing I was expecting was a practical outcome – especially one of the order of magnitude suggested by his almost casual announcement, in the last five minutes of his paper.

The Philological Society was used to papers which ended with statements of research need. Indeed, is that not how most academic papers end, with the main finding being the need for further research? And that is how Samuels seemed to be finishing. After talking about the need for more work on historical phonaesthetics – a challenge which, incidentally, has not yet been taken up – he went on:

> More urgent still is the need for descriptions of the lexical system itself.

I think many of us were expecting him to stop there, but he went on:

> Even for Modern English, no full description exists, though the gap is to some extent filled by Roget's important pioneering work of 1852, especially in its latest revision (1962). For past periods, there is nothing. To extrapolate from Roget to the *OED* may be an interesting pastime, but it can tell us only the comparative age of current forms, and cannot include words now obsolete, or words used in different meanings in the past. We need nothing less than a comprehensive historical thesaurus, with complete dates of currency, of all the forms, past and present, ever used to express single and related ideas in English, however short-lived each form – and each sense of each form – may have been.

He might have stopped there, but he goes on to underscore the significance – and the audience must have been wondering, at this point, why:

> Such a work would tell us how many and which words were available, to each writer in past periods, for the expression of a given notion (or, if you prefer, which words were either wholly or partly commutable in a given context); and it would provide the basic material necessary for detecting and solving all problems of 'semantic fields' in English [*all, note*], notably the connections, in each field, between semantic shift, verbal obsolescence and innovation. I need hardly add that it would also make a substantial contribution to literary criticism and the history of ideas.

Yes, a lovely pipedream of an idea. I seem to remember everyone nodding sagely, and then wondering whether we had heard correctly, at the next sentence:

> In an attempt to remedy this need, we have recently started a research project at Glasgow, but we are under no illusion regarding the size of the task.

And that was it. A single sentence, but it woke everyone up. Samuels gave little further detail in the ensuing question-time, but I recall a generally incredulous reaction. Several members of the Philological Society have honorary doctorates in scepticism, and my recollection of the event is that there was a general view that such a project was wishful thinking, which would founder under the weight of its own ambition. Which just goes to show that super ideas, no matter how crazy they sound, should never be written off.

Today I want to talk about another crazy idea. Sometime in the 1970s – I have lost track of the actual date – a group of lexicologists met in an Oxford pub, under the chairmanship of Laurence Urdang, the managing editor of the unabridged *Random House Dictionary of the English Language*, to discuss the desirability of what he called a 'super-dictionary' of English – a compendium of all the words known in the language. I remember Randolph Quirk was there. The motivation arose out of some exercises that had been carried out in comparative lexicography, where it had emerged that there were surprising differences in the coverage of lexical items between the unabridged general-purpose dictionaries. We expect to see differences in treatment among dictionaries – in the definitions, and in the range of information provided by an entry – indeed, that is the chief reason why dictionaries proliferate in the first place. But we don't expect general-purpose dictionaries of a similar size to have serious differences in coverage. Yet that is what we found.

I carried out such an exercise on a small sample of data for my *Cambridge Encylopedia of the English Language* (1995), comparing the unabridged *OED* and *Webster*. I took the first 57 items from letter S (the odd total simply reflects the number that would fit into the table on the page) and found that the two dictionaries had only 21 items in common – less than two-fifths. For example, *Webster* had *sabalote*, *sabal palmetto*, and *sabana*, but the *OED* didn't. The *OED* had *sabaoth*, *sabarcane*, and *sabate*, but *Webster* didn't. The *OED* has of course far more historical references and British dialect items than does *Webster*, which in turn has far more local American items. But neither covers all that there is. Reference to *Chambers* brought a cluster of items missing from both *Oxford* and *Webster* – *sabahan*, *sabbath-breach*, and *sabbath-breaker*, for example. Reference to a specialized dictionary – Willis's *Dictionary of the Flowering Plants and Ferns* – brought several others, such as *sabaudia*, *sabaudiella*, and *sabazia*. And that is just one specialist text among many. If we include abbreviations, that single SAB to SABB section would bring to light dozens more. Gale's *Acronyms, Initialisms, and Abbreviations Dictionary* had (in the 11th edition, 1987) 38 items here, including several shortened forms of general words (eg *sabbath*, *sabotage*, *soprano/alto/bass*) as well as names of organizations (eg *School of American Ballet* and *Space Applications Board*). How many of the latter we might wish to include in a superdictionary is an intriguing methodological question. American lexicographers would traditionally include encyclopedic items, and some British dictionaries these days are following suit, responding to popular demand. There are large numbers of proper names in, say, the *Longman Dictionary of Language and Culture*, where, under SAA, we will find *Saab*, *Saami*, and *Saatchi and Saatchi*. But none of these are in the *OED* or *Webster*.

Leaving proper names aside, the specialized lexicons of encyclopedic domains are not well covered in the major dictionaries, and when we reflect on the knowledge-base that is 'out there' we can see why that is. There are, apparently, some million insects already identified, with several million more awaiting description. This means that there must be at least a million lexical designations enabling English-speaking entomologists to talk about their subject. But let me generalize to the academic world as a whole. Academics in all fields are constantly innovating conceptually – that is what academics are for – and looking for new terms, or new senses of old terms, to express their new thinking. An ad hoc use of a word to express an immediate semantic need is usually called, in our subject, a *nonce-formation*. Nonce-formations are words spontaneously coined on the spur of the moment to meet an immediate communicative need. If I use a word noncely to make a point, I have no expectation that this would ever become a permanent item in English, as a neologism. Most everyday nonce-formations disappear without trace. However, it is different in the academic world, where we are constantly noncing about. It is the nature of academic enquiry to be lexically innovative. We write a paper and say at a certain point 'I shall call this X', where X may be a totally new word or an old word with a new meaning. We are, in this respect, exactly like Lewis Carroll's Humpty Dumpty: 'When I use a word ... it means just what I choose it to mean'. Absurd in a conversational context, it is the modus vivendi of academic enquiry.

Now these academic nonce-words (I called them *bonce words*, in a festschrift paper for Whitney Bolton some years ago, on the grounds that academics are supposed to have very large brains) are problematic, from the lexicographer's point of view. Unlike conversational nonce-words, they are not made on the spur of the moment; they are the product of careful thinking; and it is the intention (or at least the hope) that they should enter the (academic) lexicon as a whole, and become standard. So, what proportion of this putative academic lexicon actually finds its way into a dictionary? I have only looked at the subject I know best, linguistics, and the answer is: hardly any. In fact, I looked at just one book, *Systems of Prosodic and Paralinguistic Features in English*, where I know exactly what the bonce-formations are because Randolph Quirk and I coined them. They include general phonetic notions such as *breathiness* and *creak*, but used in a systemic way; Survey of English Usage notions in intonation such as *booster* and *prosodic subordination*; and a host of musical terms given a phonological application, such as *allegro* and *crescendo*. Of the 74 terms identified, 95 percent are not listed in the *OED* at all, and the few that are mentioned relate to a more general usage. For example, the *OED* gives only a very general notion of *paralinguistic*, and not the specifically phonological sense introduced by us. Would it be a reasonable extrapolation to suggest that perhaps 95 per cent of all academic metalanguage receives no lexicographical coverage at all? They do not necessarily get included in specialized dictionaries either, as these focus on the most widely used terms in a subject. Hardly any of the terms from the above book are included in my *Dictionary of Linguistics and Phonetics*, for example.

Non-academic specialized vocabularies also present problems, especially to do with compound words. Take this extract from a newspaper article on the health value of red wine. Red Burgundy, says the writer, 'is made with Pinot Noir, best-scoring grape for resveratrol, in a damp, mould-prone climate ... Some resveratrol is lost during barrel-ageing, and some more during long bottle-ageing. Fine Burgundy will be both barrel- and bottle-aged.' These are 'heart-friendly' wines, supporting the 'red-wine-is-best' theory. And so on. What are we to do with *mould-prone*, *best-scoring*,

and suchlike? *Mould-prone*, for example, is not an idiosyncratic usage. It may not be very frequent, but it has 5,000 hits on Google.

Under the heading of specialized lexicons, we should also include dictionaries of the nonstandard language, especially slang. The *OED* has some excellent coverage of historical slang, but it's by no means complete. As an example, here is the section on words and phrases to do with being drunk from John Ray's *Complete Collection of English Proverbs* (1670):

> He's disguised. He has got a piece of **bread and cheese** in his head. He has **drunk more than he has bled**. He has been in the sun. He has a jag or load. He has **got a dish**. He has got a **cup too much** [cf in one's cups]. He is **one and thirty**. He is dagg'd. He has **cut his leg**. He is **afflicted**. He is top-heavy. The malt is above the water. As drunk **as a wheel-barrow**. He makes **indentures** with his legs. He's well to live. He's about to cast up his **reckoning** or accompts. He has **made an example**. He is concerned. He is as drunk as **David's sow**. He has stolen a **manchet** [loaf] out of the brewer's basket. He's raddled. He is very **weary**. He drank till he **gave up his half-penny**, i.e. vomitted.

The underlined items are covered by the *OED*; the boldface items are not. There is just under 50 per cent coverage. Keeping up in such a semantic field as drunkenness is not easy. If you hear a reference to someone being *lagered*, *boxed*, *treed*, or *bladdered*, for example, you will as yet get no assistance from the *OED* – and if the *OED* hasn't got it, I doubt whether any other dictionary will.

The Ray example raises the question of the extent to which collocations and figurative expressions need to be included in a superdictionary – or, for that matter, in any dictionary. Many expressions of course are to be found in specialized collections, such as Brewer's, which, note, calls itself a ***Dictionary*** *of Phrase and Fable*. But even a brief comparison of such works with a general dictionary brings to light many discrepancies. Take similes, for example. There are several excellent collections of similes, few of which get into the mainstream dictionaries. Compare the coverage of Bartlett Jere Whiting's (1989) huge *Modern Proverbs and Proverbial Sayings* with the *OED*. I open the book randomly at p. 142 and find *as clear as crystal*, *as clean as crystal*, *as hard as crystal*, and *as white as crystal*. The *OED* has only the first, which is undoubtedly the commonest. Similarly, the *OED* has *as cool* or *cold as a cucumber*, but not *as calm as a cucumber*, which is in Whiting; and it doesn't have any *cuckoo* expressions at all, though Whiting has *as crazy as a cuckoo*, *as barmy as a cuckoo*, and *as lousy as a cuckoo*. Go back in time, as Whiting did for his earlier collection, *Proverbs, Sentences, and Proverbial Phrases from English Writings Mainly before 1500* (1968), and we find additionally *as bright as crystal* and *as sheen as crystal*.

You might be thinking that the comparisons made so far have already turned the superdictionary project into a lexicographical Mt Everest; but we ain't seen nothin' yet. For the two largest domains of lexical expansion I have still to cover. The first is the simple consequence of the rise of English as a global language. Most of the adaptation that takes place when a 'new English' emerges is in relation to vocabulary, in the form of new words (borrowings), word-formations, word-meanings, collocations, and idiomatic phrases. There are many cultural domains likely to motivate new words when English comes to be used in such places as West Africa, Singapore, India, or South Africa, and speakers find themselves adapting the language to meet fresh communicative needs. They want to talk about themselves and this

means adding to dozens of semantic fields. The country's biogeographical uniqueness will generate potentially large numbers of words for animals, fish, birds, insects, plants, trees, rocks, rivers, and so on – as well as all the issues to do with land management and interpretation, which is an especially important feature of the lifestyle of many indigenous peoples. There will be words for foodstuffs, drinks, medicines, drugs, and the practices associated with eating, health-care, disease, and death. The country's mythology and religion, and practices in astronomy and astrology, will bring forth new names for personalities, beliefs, and rituals. The country's oral and perhaps also written literature will give rise to distinctive names in sagas, poems, oratory, and folktales. There will be a body of local laws and customs, with their own terminology. The culture will have its own technology which, regardless of its primitiveness by Western standards, will have its technical terms – such as for vehicles, house-building, weapons, clothing, ornaments, and musical instruments. The whole world of leisure and the arts will have a linguistic dimension – names of dances, musical styles, games, sports – as will distinctiveness in body appearance (such as hair styles, tattoos, decoration). Virtually any aspect of social structure can generate complex naming systems – local government, family relationships, clubs and societies, and so on. Nobody has ever worked out just how much of a culture's lexicon is community-specific in this way; but it must be a very significant amount – at least 25 per cent, I would say. So, when a community adopts a new language, and starts to use it in relation to all areas of life, there is inevitably going to be a great deal of lexical creation.

When local vocabulary from all sources is collected, a regional dictionary can quickly grow to several thousand items. There are over 3,000 items recorded in the first edition of the *Dictionary of South African English* (Branford and Branford, 1978), and later editions and collections show the number to be steadily growing (there are a further 2,500 entries already added in Silva (1996)). South African Indian English alone has 1,400 (Mesthrie, 1992). *The Dictionary of New Zealand English* (Orsman, 1997) has 6,000 entries. The *Concise Australian National Dictionary* (Hughes, 1989) has 10,000. There are over 15,000 entries in the *Dictionary of Jamaican English* (Cassidy and Le Page, 1967) and 20,000 in the *Dictionary of Caribbean English Usage* (Allsopp, 1996). The small islands of Trinidad and Tobago alone produced some 8,000 (Winer, 1989), which rose to over 12,200 in Winer's later work (2009). Here are some examples of the kind of item encountered in these dictionaries – taken from the beginning of letter S in Winer (2009):

saada mahatam – type of drum rhythm
saajhe – prepare, e.g. food, house
saajhe bojhe – do something slowly – pron sadgeh bodgeh
saanay – mix together, esp foods
saar – term of address for a man's wife's younger sister
saarubhai – for various relatives, such as a wife's sister's husband
saas – another kinship term, esp for a mother-in-law

Needless to say, these don't figure in any of the front-line dictionaries.

Finally, the Internet. It is too early to say what the impact of the Internet is going to be on the English lexicon, as the phenomenon is still only a couple of decades old (the World Wide Web dates from 1991). When I compiled my *Glossary*

of Textspeak and Netspeak in 2004, I found a relatively small number of neologisms – only a thousand or so, which I considered a drop in the ocean of lexical coinage – but this was before social networking arrived. As I said, we ain't seen nothin' yet. Each technological development is going to bring lexical innovation, and if it is a linguistically attractive name this is going to generate huge amounts of new words via language play. Twitter is the perfect example, for its unusual (in English) initial consonant cluster has caught the imagination. We now have several Twitter dictionaries online, such as the Twictionary. Here are some entries:

> *actwivist, actwivism* - one who uses a Twitter message to advocate or oppose a political, social, or environmental cause
> *attwaction* - a crush on a fellow twitterer
> *attwacker* - someone who verbally assaults someone on Twitter
> *atwistocrat* - a twitter user who falsely sees self as superior
> *twatarazzi* - someone who spends all day watching celebs
> *twaddict* - someone addicted to Twitter
> *twanker* - you can guess

There are hundreds of such entries, usually with the coiner's name provided, and sometimes with an indication of popularity (click 'like').

This is the method used by the most ambitious lexicographical event ever, the Urban Dictionary, which began in 1999 and now has 6.7 million definitions of, as the Wiki entry puts it, 'slang or ethnic culture words, phrases, and phenomena not found in standard dictionaries'. Its strapline: 'Define your world.' To illustrate what happens here, I spent some time trying to find an entry that wasn't too dirty to be aired in public, and eventually I found one. The important point is to note the number of people who have said they like it or use it, and those who don't, as these figures give us an indication of the likely future of the word or sense – at least among the young-person demographic (below 25) of this site.

> *sabby*
> 1 the very best thing in the whole world. *89 up, 40 down.*
> 2 to pull a sabby, said to be specific to Nottingham, to become too drunk to actually make it out of your house. *31 up, 24 down.*
> 3 a stuck-up attitude. *49 up, 44 down.*
> 4 a short, annoying, devious creature. *24 up, 27 down.*
> 5 a way to describe an extremely stupid joke. *9 up, 9 down.*

The site attracts 15 million visitors a month, and 2,000 suggestions for inclusion a day. Lexicography has never seen anything like it.

So: any superdictionary would begin by integrating the coverage of available unabridged dictionaries, and then supplement it with reference to the lexicons of specialized domains, global varieties, and the Internet. Only a sophisticated online presence could possibly cope – and only a sophisticated management system, for it's obvious that a simple alphabetical organization would obscure rather than reveal lexical insights. There is little to be gained, other than convenience of look-up, by integrating the specialized vocabularies of different academic domains into a single list. Rather we need a thesaural approach. (I have been looking for an adjective for *thesaurus*, but have not found one. The *OED* has *thesaurial*, but that turns out to mean 'of or pertaining to the office of treasurer'. Surely the thesaurus team coined

one?) A thesaural approach, then, in which sociolinguistic dimensionality becomes a central organizing principle. It is not simply 'words in time', as the current *HTOEL* illustrates, but 'words in time and place', where 'place' includes all parameters of regional, social, and stylistic variation.

The team involved in creating a superdictionary would need expertise in taxonomy (more than anything else), computational linguistics, lexicography, semantics, sociolinguistics, and a thick skin (to deal with the online entries). Only a contributor-based model would be practicable, along the lines of the 200-strong team I brought together to compile *The Cambridge Encyclopedia*. And a good life expectancy. I am reminded of Dr Strong in Charles Dickens's *David Copperfield*. In Chapter 16 the young David arrives at Dr Strong's school. Dr Strong is described as working on a 'a new Dictionary [of Greek roots] which he had in contemplation'. David tells the story. 'Adams, our head-boy, who had a turn for mathematics, had made a calculation, I was informed, of the time this Dictionary would take in completing, on the Doctor's plan, and at the Doctor's rate of going. He considered that it might be done in one thousand six hundred and forty-nine years, counting from the Doctor's last, or sixty-second birthday.' I do not know how long it would take to compile a superdictionary. Probably as long as it took to compile the first edition of the *OED*, or the *Historical Thesaurus*. The first time-scale didn't put off James Murray. The second didn't put off Michael Samuels. I hope someone will find this project just as appealing.

References

Allsopp, Richard. 1996. *Dictionary of Caribbean English Usage.* Oxford University Press.

Ayto, John (ed.). 2005. *Brewer's Dictionary of Phrase and Fable*, 17th edition. London: Weidenfeld & Nicolson.

Branford, Jean and William Branford. 1978. *Dictionary of South African English*. Cape Town: Oxford University Press.

Carroll, Lewis. 1872. *Through the Looking Glass*. London: Macmillan.

Cassidy, F. G. and R. B. Le Page. 1967. *Dictionary of Jamaican English*. Cambridge University Press.

Crystal, David. 1995. *The Cambridge Encylopedia of the English Language.* Cambridge University Press.

2000. Investigating Nonceness: Lexical Innovation and Lexicographic Coverage. In Robert Boenig and Kathleen Davis (eds.), *Manuscript, Narrative and Lexicon: Essays on Literary and Cultural Transmission in Honor of Whitney F. Bolton* (Lewisburg: Bucknell University Press; London: Associated University Presses), 218–29.

2004. *A Glossary of Netspeak and Textspeak*. Edinburgh University Press.

2008. *Dictionary of Linguistics and Phonetics*, 6th edition. Oxford: Blackwell.

Crystal, David and Randolph Quirk. 1964. *Systems of Prosodic and Paralinguistic Features in English*. The Hague: Mouton.

Gove, P. B. 1961. *Webster's Third New International Dictionary*. Springfield: Merriam.

Hughes, Joan (ed.). 1989. The *Concise Australian National Dictionary*. Melbourne: Oxford University Press.

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

Mesthrie, Rajend. 1992. *Lexicon of South African English*. Leeds: Peepal Tree Press.
Orsman, Harry W. 1997. *The Dictionary of New Zealand English*. Auckland: Oxford University Press.
Ray, John. 1670. *Complete Collection of English Proverbs*. London: Allman.
Robinson, Mairi (ed.). *Chambers' 21st Century Dictionary*. 1996. Edinburgh: Chambers.
Samuels, Michael. 1965. The Role of Functional Selection in the History of English. *Transactions of the Philological Society*, 15–40.
Silva, Penny (ed.). 1996. *A Dictionary of South African English on Historical Principles.* Oxford University Press.
Simpson, John and Edmund Weiner (eds.). 1989. *Oxford English Dictionary*, 2nd edition. Oxford University Press.
Summers, Della (ed.). 1992. *Longman Dictionary of Language and Culture*. Harlow: Longman.
Stein, Jess and Urdang, Laurence (eds.). 1967. *Random House Dictionary of the English Language.* New York: Random House.
Towell, Julie E. and Helen E. Sheppard (eds). 1987. *Acronyms, Initialisms, and Abbreviations, Dictionary*, 11th edition. Detroit: Gale.
Whiting, Bartlett Jere. 1989. *Modern Proverbs and Proverbial Sayings.* Cambridge, MA: Harvard University Press.
Whiting, Bartlett Jere and Helen Wescott Whiting. 1968. *Proverbs, Sentences, and Proverbial Phrases from English Writings Mainly before 1500.* Cambridge, MA: Belknap Press.
Willis, J. C. 1985. A *Dictionary of the Flowering Plants and Ferns*, 8th edition. Cambridge University Press.
Winer, Lise. 1989. Trinbagonian. *English Today* 18, 17–22.
Winer, Lise. 2009. *Dictionary of the English/Creole of Trinidad and Tobago*. Montreal and Kingston: McGill-Queen's University Press.