Good Old-fashioned Lexicography: Human Judgment and the Limits of Automation

Michael Rundell

(The Lexicography MasterClass and Honorary Fellow of the Information Technology Research Institute, University of Brighton, U.K.)

Abstract

Some commentators have suggested that, as corpora and data-mining software improve, the role of the human lexicographer may gradually become less central, perhaps reaching a point where the lexicographer's main function is to 'package' a linguistic description that has been arrived at primarily through the interaction of smart software and high-quality data. This paper takes an almost opposite view, and argues that the vast amounts of information now available to us – while a wonderful resource that will underpin much more reliable dictionaries – will call for even more highly-skilled and linguistically-aware editors, whose role in *interpreting* the data and *synthesising* dictionary text from it becomes ever more demanding.

1. What are lexicographers for?

> 'In lexicography, as in other arts, naked science is too delicate for the purposes of life'
>
> Samuel Johnson, *The Plan of a Dictionary* 1747. 4

The year 2001 is an appropriate moment for us to be thinking about the power and capability of computers. Arthur C. Clarke's futuristic novel, *2001– A Space Odyssey*, imagined a machine that could hold conversations and think for itself – or at least simulate this process convincingly enough to give the impression of being truly intelligent. Within the broader debate about the potential for computers to learn, exercise judgment, and even have consciousness – which is often associated with Alan Turing's famous paper (Turing 1950), where he poses the question 'Can machines think?' – there is a more localized issue of special interest to those of us who write dictionaries. In essence, it is this: given the ever-increasing capacity of computers to store vast amounts of linguistic data, coupled with the growing sophistication of the software tools available for analyzing this data, can we now foresee a time when human beings will play only a subordinate, organizing role in the process of producing descriptions of languages?

The question has been asked before. Gregory Grefenstette, for example, poses the question 'Will there be lexicographers in the year 3000?' (Grefenstette 1998), and outlines a series of software routines, all of them already possible and most now in regular use, which go a long way towards automating the process of linguistic data analysis. Grefenstette concedes that 'so long as [dictionaries] are printed we will need the reasoned condensations that only lexicographers provide' (ibid. 39) But the implication is that for *online* reference works, where there are fewer constraints on the size of the data store, the human contribution may indeed become rather marginal – hardly an encouraging outlook for the professional (human) linguist or lexicographer. As far back as 1987, John Sinclair claimed that 'a fully automatic dictionary is [now] at the design stage' (Sinclair 1987. 152), which I take to mean a dictionary whose description of language is achieved primarily through the interaction of intelligent software and large corpora, with minimal intervention by humans. In this model – I am making some assumptions here since Sinclair's paper does not go into detail about the 'automatic dictionary' – the computer would analyze huge volumes of corpus data and thus arrive at reliable generalizations about those linguistic features (meanings, syntactic behaviour, collocational preferences, register-specific uses, and so on) that appeared to be most typical of the language being studied[1]. One of the attractions of this model (to its proponents, at least) is that it eliminates from the process the exercise of human intuitions about language, which, we are frequently told, are too partial and too subjective to be a reliable guide to the way people really speak and write. What is envisaged here, then, is a scenario in which the increasing power and sophistication of machines goes hand in hand with a corresponding reduction in the role of the human editor.

Much of this looks very plausible and might also (if they were to hear about it) be as attractive to publishing managers as the idea of virtual actors is to Hollywood producers. My argument in this paper, however, is that the growing contribution of computers to the lexicographic process will entail not the progressive 'de-skilling' of lexicographers but – paradoxically, perhaps – an even greater need for skilled human editors with a good grounding in relevant linguistic disciplines *and* highly developed intuitions about language. In other words, for the foreseeable future there will still be a demand for ordinary

[1] See also Barnbrook (1996: 136), who puts a little more flesh on the bones of this idea: 'The provision of NLP capability within the basic lexicographic tools could ... assist in the production of definitions and the selection of suitable example texts. Ultimately ... any changes in the behaviour of words could automatically be detected and assessed by the software.'

lexicographers – and of course an even greater demand for extraordinary ones like Sue Atkins.

2. The corpus revolution

The technical developments that have transformed lexicography over the past 20 years need little elucidation here, and are in any case described elsewhere in this volume (see especially Kilgarriff and Tugwell). Writing in a recent edition of the *EL Gazette* (the trade journal of the English language teaching profession), ELT writer Michael Lewis made the incontrovertible point that 'the first Cobuild project changed the face of dictionary-making'. It did so by establishing a new paradigm in which corpora of naturally-occurring text would provide the primary data source for all good dictionaries. Though pockets of resistance remain (notably among some of the major U.S. dictionary publishers), corpus lexicography is regarded as a given within my own subfield of pedagogical dictionaries for learners of English as a Second Language. (In the rest of this paper, references to dictionaries and dictionary-making will generally relate to this branch of lexicography.) The small, 7-million-word corpus of the early Cobuild years, with its static concordance printouts generated in a one-off operation by an industrial-strength mainframe, has given way to a situation where hundreds of millions of words of text can be stored, and queried in real time in a variety of ways, on any inexpensive personal computer. Against this background, the focus in corpus lexicography has begun to move away from issues such as the size and composition of corpora (which preoccupied us in the 1980s and 1990s) towards the newer challenges of how best to extract lexicographically relevant information from very large text databases. For lexicographers, too, have to deal with their own particular flavour of that besetting problem of contemporary life, information overload.

To put this in perspective: a dictionary-writer working with a 200-million-word corpus would have access to around 1500 concordance lines for a medium-frequency word like **forge**, 3500 lines for **forgive**, and 25,000 lines for **forget**. No human editor, even without the time constraints that inevitably apply to most dictionary projects, could make sense of this much data by scanning it in 'traditional' concordance format: hence the need for some form of automated summarization. Though Church and Hanks' famous 1989 paper turned out to be something of a false dawn for lexicographers, it pointed the way to a new generation of 'lexical profiling' software that would analyze large corpora and produce statistical summaries of considerable delicacy.

Probably the best current example of this genre is Kilgarriff and Tugwell's 'Word Sketch' software, whose features and implementation are fully explained elsewhere in this volume. It may, however, be worth adding a brief evaluation of its practical utility in a real lexicographic project, now that a complete new dictionary (Rundell 2002) has been created by compilers who had access both to conventional concordancing software and to Word Sketches for the core vocabulary of English. The original intention was that the Word Sketches would supplement existing resources, specifically by enhancing and streamlining the process of identifying salient collocates of various types (such as operating verbs or intensifying adverbials) for a given dictionary headword. It quickly became clear, however, that for most editors the Word Sketches came to be the preferred starting point for looking at a word. What appeared at first to be a set of discrete lists, each focussing on a specific combinatorial frame, turned out in practice to be more than the sum of its parts. For the Word Sketches, by encapsulating the key features of a word's behaviour, provide editors with a compact and revealing snapshot which contributes powerfully to the identification of word meanings (one of the hardest of all lexicographic tasks). Recent experience suggests, therefore, that lexical profiling software of this type may have quite significant methodological implications for the practice of lexicography (see now Kilgarriff & Rundell 2002).

With massive volumes of text now at our disposal, and even more sophisticated data-mining tools already under development (including the next incarnation of the Word Sketch software), it may appear that we are progressing steadily towards ever-greater automation of the dictionary-making process. Could it be that those who insist on a continuing, and central, role for human language-analyzers are simply guilty of what Turing called the 'Heads in the Sand' tendency – a refusal to contemplate that machines can do our thinking for us because (in Turing's caricature) 'the consequences of machines thinking would be too dreadful, [so] let us hope and believe that they cannot do so' (Turing 1950. 444)? When computer power automates processes that once involved enormous human effort, it is not surprising that there is some resistance among those who have invested so much of their time and effort in the manual process.

The counter-argument, however, is that this 'linear' view of recent progress may misrepresent the reality. An alternative (and for me, more persuasive) interpretation is to see the process as cyclical rather than linear. According to this view, a corpus-driven approach to lexicography enables us to achieve a more reliable and more complete language description, and helps us to resolve

many of the problems that we were already grappling with. But in the process, it uncovers entirely new and unsuspected layers of complexity. Michael Stubbs recently made the point that 'Corpus linguistics provides quantities of data which were inconceivable a few years ago, so that it is not surprising that these data are now causing problems of *interpretation*' (Stubbs 2001. 169, emphasis mine). This is the territory we have now entered – a probabilistic world where we discern 'tendencies' or 'norms' or what Patrick Hanks has often called 'preferences' (e.g. Hanks 2000a. 6). The problems we have solved so far in corpus lexicography are – it now appears in hindsight – of a relatively straightforward type, mainly in the realms of observable 'fact'. And, crucially, they relate quite largely to already familiar linguistic categories. Progress here has, without question, been impressive, and the very real benefits for dictionary-users should not be underestimated. (Monolingual dictionaries for learners of English have been transformed almost beyond recognition.) But this does not necessarily bring us closer to complete understanding (whatever that means). Rather, the process seems to be recursive: familiar problems get solved, and at the same time completely new ways of interpreting the data arise. (Which of course is what makes corpus lexicography such an addictive occupation.)

3. Interpreting data (1): linguistics and lexicography

So far we have mainly discussed the provision of data – in ever greater volumes and higher quality, and with increasingly smart software tools to facilitate its analysis. But Stubbs' point about interpretation reminds us that linguistic data is merely the starting point. If data is the input, and dictionaries are the output, then – as Atkins shows with characteristic clarity – getting from one to the other entails two distinct stages, which collectively represent the process of corpus lexicography: analysis and synthesis (Atkins 1993. 7-8). Analysis, or the 'interpretive' stage of dictionary-making, involves a bottom-up process whereby we attempt to discern and abstract the underlying order and regularity from what sometimes seems like the chaos of disparate individual instances of words in use. Alongside the increasingly significant contribution of computational linguistics to this task (see previous section) there is an important role too for theoretical linguistics, in helping lexicographers to develop frameworks that will guide this organizing process. Paradoxically, the most valuable insights here tend to come not from so-called metalexicographers (whose influence on lexicographic practice may be less profound than is sometimes imagined) but from linguists working within their own fields who 'do not tell lexicographers

what to do ... [but] show us different ways of looking at language and word meaning, which we can take and adapt to our needs' (Atkins 1993. 29).

Fields such as lexical semantics, prototype theory, pragmatics, and of course Sue Atkins' own specialization, frame semantics, have already contributed very significantly to the way lexicographers analyze language. A good recent example of the interaction between linguistic theory and lexicographic practice is the whole field of phraseology and the combinatory tendencies of words. An enormous amount of work has been done in this area over the past 20 years or so, much of it exploiting, and made possible by, the new availability of large corpora (for a good recent overview, see Cowie 1998). Consequently, we now have a far better understanding of phraseology, and especially of the central role of prefabricated units of language in the way that we store, process, and articulate language. This research has filtered right through into language-teaching practice, through classroom-oriented books like Michael Lewis's *The Lexical Approach* (Language Teaching Publications, 1993). And of course, it has direct relevance to dictionary-making. Following the theoretical lead, dictionaries (especially those aimed at language learners) have moved towards a more phrasally-oriented approach, with greater emphasis on multiword units of various types (see e.g. Rundell 1998. 320).

In this context, it is worth looking briefly at two other fields that have not yet (as far as I know) had much impact on practical lexicography, but which may have much to offer.

3.1. Semantic prosody

The concept of semantic prosody originally surfaced in the work of Bill Louw (Louw 1993), and was taken up by linguists such as John Sinclair, Michael Stubbs, and Michael Hoey. It describes the way that aspects of a word's meaning are often present in the surrounding text. While *collocation* describes the tendency of Word A to associate regularly with Word B, semantic prosody characterizes the way that a *whole semantic class* may have a strong tendency to be associated with a given word. Thus, in a well-known example, Stubbs shows how the verb **cause** has a 'strongly negative semantic prosody' (Stubbs 1996. 173): while the verb essentially means 'to make something happen', corpus data shows us that there is a very marked preference for it to be followed by a 'negative' object, such as **disruption, disease, death**, or **confusion**. Similarly Hoey, investigating the related word **consequence**, shows that the ratio of negative to positive adjectives modifying **consequence** is around 5:1, while for **result** it is 2:5. This leads to the observation that, in broad terms, **consequence**

has a clearly negative semantic prosody, and **result** a mainly positive one. The lexicographic significance of this is pointed up by Hoey's conclusion that 'only looking at the individual words [that is, at specific collocates] disguises a more powerful generalization' (Hoey, in preparation) – and powerful generalizations are, after all, precisely what lexicographers aim to discover and convey to dictionary users.

Semantic prosody, in concordance-scanning terms, involves looking a little further beyond the node word than lexicographers have become accustomed to doing. Consider, for example, the frame:

		the television
		the TV
verb phrase	**+ in front of**	**+ a game show**
		EastEnders
		etc.

Corpus data shows that while people sit *at* their computer screens (implying purposeful interaction), they sit, sprawl, plonk themselves, vegetate, or curl up *in front of* their televisions (implying mindless passivity). Over and above all these frequently occurring verbs, the collocate of choice in these circumstances is, very clearly, the word **slump**:

```
heir peers back in Britain would be slumped in front of the telly as the g
 in search of mass entertainment or slump in front of the television, seek
 leading a sedentary life,  sitting slumped in front of the television,los
r instruments, and our leisure time slumped in front of a television set,
re readily than a couch potato. But slumping in front of the television se
for advice,while their husbands are slumped in front of the tenth re-run o
de cheap entertainment for a public slumped in front of the telly.
```

We may not yet have worked out, in every case, how to reflect insights like this in dictionary text, but they clearly support Stubbs' more generally revealing observation that 'a major finding of corpus linguistics is that pragmatic meanings, including evaluative connotations, are more frequently conventionally encoded than is often realized' (Stubbs 2001. 153). All of which presents interesting challenges for the lexicographic community.

3.2. Metaphor

Though Lakoff and Johnson's seminal text on this subject (Lakoff & Johnson 1980) is over 20 years old, the ramifications of their work are only now beginning to percolate down into practical language-teaching materials[2]. The pedagogical potential of this view of metaphor is enormous, for at least two

[2] See for example Wright 1999, which draws heavily on *Metaphors We Live By*.

reasons. First, it helps the language learner to understand underlying language systems and perceive links between the formally unrelated lexical items through which particular concepts are lexicalized[3]. Secondly, an understanding of the most pervasive metaphors in a target language will facilitate both the decoding and the *retention* of previously unencountered vocabulary (see now Boers 2000, reporting a recent experiment). The lexicographic implications of this – at least for those of us involved in producing learner's dictionaries – are just beginning to be exploited. To give one example: corpus data for the word **conversation** includes lines like the following:

```
e music while they ate and the conversation moved from the complexitie
s were devoted to bringing the conversation round to the topic of food
 village. </p> One evening the conversation turned to commando raids d
ized he was trying to lead the conversation away from her husband,  to
that she was able to steer the conversation in the direction that, how
Desperately, Celia steered the conversation round to her and Brian 's
gether easily,  casually,  the conversation  drifting.  Jane had shown
returned with the drinks,  the conversation returned to the antics per
&equo </p> From knitting,  the conversation  moved,  via dressmaking,
s later the fun dried up.  The conversation started to take a vaguely
the suddenly intimate turn the conversation had taken. </p> " Nothing
 silence,  wondering where the conversation was leading. </p> &quot I
erage Cagliaritano without the conversation drifting round to the camp
```

Word Sketch data for **conversation** points in the same direction, with the list of verbs where **conversation** is the subject including items such as **drift, revolve around, veer, wander**, and **move**. All of this prompts a Lakoff-and-Johnson-style hypothesis that 'A CONVERSATION IS A JOURNEY' (cf. Lakoff & Johnson 1980. 91ff.), and sure enough, a wide range of other lexical items confirm that a metaphor of this type underlies a great deal of conversation-related vocabulary in English. Consider, for example, expressions like these:

I can't quite see where this argument **is heading**
It was a useful meeting – we **covered a lot of ground**
I feel you're **on the wrong track** here
We eventually **arrived at** a conclusion
The discussion **drifted** rather aimlessly
I think you've **wandered off** the topic here
We kept **going round and round** in circles

Material of this type has now been introduced into a learner's dictionary (Rundell 2002), in the form of usage notes showing how 40 or so of the

[3] For example, the various words and phrases encoding the notion of 'anger': Lakoff 1987: 380ff.

commonest metaphors in English are typically encoded in common words and phrases. This is a modest beginning, but another good example of the interaction of linguistic theory, corpus data, and lexicography.

In all these cases, we see an iterative process at work: one set of data gives rise to a useful theoretical generalization, which in turn helps lexicographers and linguists to discern patterns and systems operating across much larger stretches of text. At the very least, this contributes to the interpretive stage of the dictionary-making process, and in many cases the resulting insights are reflected in actual dictionary text.

4. Interpreting data (2): things can only get fuzzier

Among the many amazing revelations of corpus linguistics, none is more striking than the recognition that almost every linguistic category one can think of is at best a prototype. Long before the corpus revolution we were familiar with the notion that objects cannot always be assigned unambiguously to watertight categories, and instead we should think in terms of 'degrees of category membership'. But just as there are prototypical (and less prototypical) birds and cups, similar boundary problems arise with familiar linguistic categories. Criteria for describing text-types, for example, have traditionally included the attribute 'mode of discourse', which used to be seen as a binary choice between spoken and written text. But email messages (especially when written by skilled keyboarders who type almost as fast as they speak) exhibit many of the features of spontaneous conversation, and thus straddle conventional boundaries. (The even newer form, text-messaging, is still harder to categorize.) Similarly with word classes. The basic categories are serviceable enough in most cases (even if the term *adverb* is a repository for a rather alarming range of functions), but there are plenty of exceptions. When automated taggers assign 'portmanteau tags' such as AJ0-VVG or NN1-AJ0, it is not necessarily a sign of inadequacy in the programs: rather, their uncertainty mirrors a genuine (and probably intractable) lack of clarity, or at least of consensus, among human analysts. At what point, for example, does **forgiving** change from verb to adjective? Or, in expressions like *summer vegetables* and *City's summer interest in Middlesborough midfielder Phil Stamp*, is **summer** a noun or an adjective? An even more adjectival noun is **core**: management gurus forever talk about core values, core competencies, and core business activities, and in contexts like this **core** is almost always used to modify other nouns. But there are signs now of it finally crossing the species barrier (following a route already taken by the word **key**) into true adjectival territory:

I don't think there will be a lot of people buying big mainframes, but they are **core** to the business for the people who have them.
Core to the design is the provision of three rows of seats with places for seven adults.

The point is not that traditional word classes are suddenly exposed as faulty categories, but simply that we cannot regard them as watertight groupings to which items can be assigned with absolute confidence and with no 'leakage'.

Which brings us, inevitably, to the fuzziest category of all, that of word meaning. There is general agreement that the same word-form can mean different things in different contexts. But this unexceptionable premise is a very long way from the notion that, for any given word, there is a well-established and generally agreed inventory of distinct and mutually-exclusive senses. Here again, the endemic fuzziness has long been recognized[4], but access to large corpora has greatly sharpened the perception that word meaning can be regarded as (at best) yet another form of prototype. There is an interesting parallel here with the classification of species in the natural world. In his recent update of *The Origin of Species*, the geneticist Steve Jones (Jones 1999) shows how access to more detailed information – at the genetic level – has forced scientists to re-assess the discrete categories established by Linnaeus 250 years ago. In reality, the DNA evidence suggests that 'species can, in the new world of molecules, no longer be seen as absolutes'. They are not so much distinct units as rough groupings of individuals, each with its own unique attributes. Jones concludes that 'Whatever species may be ... they are not fixed. Instead, their boundaries change before our eyes ... differences blend into one another in an insensible series'.

It would be difficult to find a better description of how word meaning works. The goal of automated word sense disambiguation (of fundamental importance to NLP) is confronted by increasing doubts among lexicographers – fuelled by large quantities of corpus data – as to whether there is anything there to disambiguate. Sue Atkins' own position on this issue ('I don't believe in word senses') has passed into lexicographic folklore, while the subject has also been much discussed by Patrick Hanks. Hanks (2000b.) proposes a model where a word does not have separate meanings but rather a set of meaning potentials, each of which may be activated in particular contexts. This view of meaning presents an interesting challenge for lexicography, because the way that

[4] For example by Apresjan (1974): 'Explanatory dictionaries greatly exaggerate the measure of discreteness of meanings and are inclined to set clear-cut borders where a closer examination...reveals only a vague intermediate area of overlapping meanings' (ibid. 9)

dictionaries conventionally handle meaning divisions – with a 'flat' structure consisting of individual numbered senses – does not reflect reality: it 'creates a false picture of what really happens when language is used' (Hanks 2000b. 205). Two recent attempts to resolve this problem are worth a brief look. In the *New Oxford Dictionary of English* (*NODE*, 1998), itself a Hanks brainchild, entries are divided into one or more 'core' senses, each of which 'acts as a gateway to other, related subsenses' (Introduction). Thus for example the entry for the verb **escape** begins with a general definition of the 'core' meaning ('break free from confinement or control'), which is followed by several subsenses describing more specialized uses (such as gas or liquid escaping from a pipe). A variation on this approach is used in the *Macmillan English Dictionary for Advanced Learners* (*MED*, 2002). Here, the entry structure reflects a view that, in many cases, a word will have *some* clearly distinct meanings that conform quite well to the conventional dictionary model, but then other much fuzzier *meaning-clusters*, where a basic semantic core is elaborated, in real text, in a variety of ways. In practical terms, this means that the entry for **escape** (unlike *NODE*'s) accords full meaning status to ideas such as accidental leakage from a container (*Five tonnes of crude oil had escaped into the sea*) or failure to remember or recognize something (*It had not escaped her attention that he was late again*), but treats the 'getting away from dangerous or unpleasant situations' idea as a meaning-cluster with several subsenses. Both approaches allow us to show the underlying relatedness among the 'meanings' of essentially monosemous (or at least, oligosemous) words like **escape**, and also make it easier to account for semantic nuance, speaker attitude, and metaphor (see Appendix for both these entries).

Neither policy works perfectly in every case, but both are a move in the right direction: they recognize 'fuzziness' and attempt to create lexicographic structures that reflect it. The process we see here is one that begins with a theoretical observation, which is then corroborated by corpus data and which, finally, drives an effort to achieve a more linguistically plausible and (for dictionary users) more intuitively satisfying account of word meaning.

5. Good old-fashioned lexicography

This process forms part of what Sue Atkins has called 'synthesis' (Atkins 1993. 7ff.) – the point at which analyzed corpus data is turned into publishable dictionary text. Much of this relates to the way linguistic features are described and presented (including, inter alia, strategies for handling polysemy, as discussed above). There is also, however, the issue of *selection*, and here Sue's

notion of 'lexicographic relevance'[5] (like all the best insights, blindingly obvious after a moment's reflection) has immense value. Consider, for example, the following entries from well-known learner's dictionaries:

(1)

in·ar·ti·cu·late /£ [UK phonetics], $ [US phonetics] / *adj* unable to express feelings or ideas clearly, or expressed in a way that is difficult to understand • *When it comes to expressing their emotions, most men are hopelessly inarticulate.* • *His speech was inarticulate and it was obvious he had been drinking.*
in·ar·ti·cu·late·ly /£ [UK phonetics], $ [US phonetics] / *adv* • *I'm afraid I'm expressing myself rather inarticulately.*
in·ar·ti·cu·la·cy /£ [UK phonetics], $ [US phonetics] / ,
in·ar·ti·cu·late·ness /£ [UK phonetics], $ [US phonetics] / *n* [U] • *The inarticulacy of most politicians makes me wonder how they ever managed to get themselves elected.*

(2)

Norwegian [phonetics] (**Norwegians**)
1 Norwegian means belonging to or relating to Norway, or to its people, language, or culture: *The main road from Murmansk to the Norwegian border is still closed to foreigners ...I stood there breathing the fresh Norwegian air.* • A **Norwegian** is a person who comes from Norway: *Many Norwegians feel that Norway is a culturally young country.*
2 Norwegian is the language spoken by the people who live in Norway: *It is interesting that Grainger spoke Norwegian.*

Neither entry could be criticized as being 'wrong' (in the sense of conveying false information), nor is the presentation noticeably at fault: the definitions and examples are clear enough and unlikely to pose problems for advanced learners of English. And yet... both entries, in my view, fall down badly in terms of relevance. In the case of (1), the enormous amount of space devoted to derived forms is of questionable value for the intended user. The two nominalized forms, **inarticulacy** and **inarticulateness**, for example, appear a total of *seven*

[5] For example: 'During the synthesis stage, the compiler extracts from the collection of ordered facts those that are *relevant to the particular dictionary being written*' (Atkins 1993. 8, emphasis mine). See also Fillmore & Atkins 1998.

times in the 100-million-word BNC (and the derived adverb, just three times): what, realistically, is the likelihood of the average advanced learner ever encountering these extremely rare words? Is it great enough to justify giving two examples of their use? Entry (2) raises similar issues: **Norwegian** certainly rates an entry because it is not a predictable derived form (in the way that **Bulgarian**, for example, is). But do advanced learners really need to be told how this noun is pluralized, and do they really need all these examples of the word in use? What is their function? In the vast majority of cases, the user will look up this word having seen it in context and been unsure of its meaning: for this reference purpose, a simple entry showing that **Norwegian** is related to **Norway** will be more than adequate. Users of pedagogical dictionaries need examples either to help elaborate the meaning of concepts that are difficult to describe, or to serve as models for production, especially in the case of words whose combinatorial behaviour is complex and (to learners) unpredictable. What characterizes both these entries is a failure to distinguish information that is merely *true*, from information that is *relevant*.

These questions are far from trivial because, in a paper dictionary, any space used for showing one information-category is, necessarily, no longer available for any other use. A lack of real clarity about the issue of relevance will thus have very significant implications for the degree to which a dictionary can answer the multifarious questions that its users will ask of it. As Johnson lugubriously recognized, 'they that take a dictionary into their hands have been accustomed to expect from it, a solution of almost every difficulty' (Johnson 1747. 5) – in other words, dictionary users want it all. Our job as lexicographers is not to attempt the impossible task of catering for every conceivable need, but to develop an informed, 'utilitarian' view (in the Jeremy Bentham sense) of which precise subset of all the available information is relevant to the needs of the greatest number of users in the greatest number of situations.

Before returning finally to the question of the respective roles of computers and human editors, let us look at some data for a word that raises a typical cross-section of the problems that lexicographers face – and find ways of resolving – on a daily basis.

Like many adjectives, **old-fashioned** seems to draw much of its meaning (in the broadest sense of the term) from its context. Disambiguation is by no means straightforward. We find, for example, a range of speaker-attitudes, going all the way from negative, through neutral, to very positive:

'I'm ready now, darling, I'll just put my scarf on.' Sousan looked pained. 'No one wears headscarves in London, Mummy. It's very **old-fashioned**.'

Fortunately, the number of these **old-fashioned** classes seems to be gradually falling.

She had to hire someone to clean her house. You can bet her **old-fashioned** husband won't offer to do half!

Paula had no patience for making conversation with Gran, who tended to have very **old-fashioned**, dyed-in-the-wool ideas.

Here, old-fashionedness has connotations of outdatedness and irrelevance: it arouses irritation or disapproval. In many cases, though, the word is used as a value-free descriptive adjective, especially of everyday objects and furnishings:

At the back of the house ... the big, **old-fashioned** bathroom with its noisy pipes and its huge wood-surrounded bath.

On its top was a simple oak cross ... and an **old fashioned** black telephone, the receiver off the rest and lying on its side.

Patrons recline in an **old-fashioned** barber chair...

Hilbert and Lewis and Beryl sat in **old-fashioned** deck chairs with striped canvas seats.

Finally, there are many situations where being old-fashioned is seen as a virtue – evoking a sense of nostalgia for 'the good old days':

We're a very small, **old-fashioned** type of club [with the subtext: 'And that's the way we like it.']

The reception area [is]... decorated to conform to the same image, conveying an image of discreet, **old-fashioned** comfort and luxury.

The real way to improve the health of the capital city 's people lies with such **old-fashioned** concepts as full employment, decent housing and good education.

The story...resulted from '**old-fashioned** gumshoe reporting'

Whatever happened to good **old-fashioned** values?

The word's chameleon-like quality sometimes leads speakers to be explicit about which attitude they are invoking in given context:

He is, in the best sense, an **old-fashioned** doctor.

Something even more interesting happens when we narrow our search to the expression **good old-fashioned**. There are of course plenty of corpus instances showing the (expected) positive sense:

Mine's [=my watch] a **good old-fashioned** proper mechanical wind-up job.

Domestic security is simply a matter of **good old-fashioned** common sense.

It demonstrates that **good old-fashioned** methods can mean tastier meat.

It's a miracle of **good old-fashioned** craftsmanship.

But there are almost as many cases where the effect is quite different:

businesses that thrive on paranoia .. and **good old-fashioned** nosiness

There remain **good old-fashioned** nationalist dictators ...

...the robberies and the shootings – you know, **good old-fashioned** New York City crimes

How had a **good old-fashioned** food scare mutated into serious political trouble?

...bears a suspicious resemblance to **good old-fashioned** idle speculation

We also hear about good old-fashioned guilt, greed, and self-interest. Exactly what is happening here is a matter for interpretation, but there is an observable tendency to use this expression in an ironic way, when talking about things which – though regrettable – are also somehow comfortingly familiar. Bill Louw comments on the way that irony 'relies on a collocative clash' (Louw 1993.157). The effect, in other words, depends on the disjunction that arises when expected collocates (like 'common sense' or 'patriotism') are replaced by something less edifying.

And so to the question that we started with. The process of taking data like this and turning it into dictionary text that is both appropriate and accessible to a specific user-group is one of some complexity. Accounting for the behaviour of a word like **old-fashioned** raises issues of word sense disambiguation (how many meanings are there here? Only one, according to some dictionaries, several according to others) and might well be informed by insights from – among other fields – lexical semantics, pragmatics, and semantic prosody. And all this might constitute just one quarter of an average day's effort for the working lexicographer. The wonderful thing about technology is that it can supply us with the volume of data that we need (and, increasingly, with the software for summarizing its salient features) in order to uncover and describe linguistic behaviour of this type. But the idea that the interpretive and 'synthetic' parts of lexicography can be automated to any significant degree seems to me unlikely and possibly misguided. For the forseeable future, tasks like this will be most effectively performed by a collaborative partnership of humans and machines. For we require not only high-quality data and cutting-edge software, but also that rare combination of editorial judgment, market knowledge,

linguistic awareness, and good old-fashioned intuition that Sue Atkins possesses in such abundance.

Postscript

> 'We can only see a short distance ahead, but we can see plenty there that needs to be done.'
>
> (Turing 1950.460)

References

Apresjan, Ju. D. 1974. 'Regular Polysemy'. *Linguistics* 142: 5-32.

Atkins, B.T.S. 1993. 'Theoretical Lexicography and its Relation to Dictionary-making'. *Dictionaries* (Journal of the DSNA) 14: 4-43.

Barnbrook, Geoff 1996. *Language and Computers.* Edinburgh: Edinburgh University Press

Boers, Frank 2000. 'Metaphor awareness and vocabulary retention'. *Applied Linguistics* 21/4: 553-571.

Church, Kenneth & Patrick Hanks 1990. 'Word association norms, mutual information, and lexicography'. *Computational Linguistics* 16: 22-29.

Cowie, A.P. (ed.) 1998. *Phraseology: Theory, Analysis, and Applications.* Oxford: Clarendon Press.

Fillmore, Charles J. & B.T.S. Atkins 1998. 'FrameNet and Lexicographic Relevance'. *Proceedings of the First International Conference On Language Resources And Evaluation*, Granada, Spain, 28-30 May 1998.

Grefenstette, Gregory 1998. "The future of linguistics and lexicographers: will there be lexicographers in the year 3000?'. Thierry Fontenelle et al. (eds.) *EURALEX 1998 Proceedings.* Liège University of Liège: 25-41.

Hanks, Patrick 2000a. 'Contributions of Lexicography and Corpus Linguistics to a Theory of Language Performance'. *EURALEX 2000 Proceedings.* Stuttgart: University of Stuttgart: 3-13.

Hanks, Patrick 2000b. 'Do word meanings exist?'. *Computers and the Humanities* 34: 205-215.

Hoey, Michael in preparation. 'Beyond collocation'

Johnson, Samuel 1747. *The Plan of a Dictionary.*

Jones, Steve 1999. *Almost like a Whale: The Origin of Species Updated.* London: Doubleday.

Kilgarriff, Adam & Michael Rundell 2002. 'Lexical profiling software and its lexicographic applications: a case study'. A. Braasch et al. (eds.) *EURALEX 2002 Proceedings.* Copenhagen: University of Copenhagen.

Lakoff, George & Mark Johnson 1980. *Metaphors We Live By.* Chicago: Chicago University Press.

Lakoff, George 1987. *Women, Fire, and Dangerous Things.* Chicago: Chicago University Press.

Louw, Bill 1993. 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies'. M. Baker et al. (eds.) *Text and Technology.* Amsterdam: Benjamins: 157-176.

Rundell, Michael 1998. 'Recent trends in English pedagogical lexicography'. *International Journal of Lexicography* 11.4: 315-342.

Rundell, Michael (Editor) 2002. *The Macmillan English Dictionary for Advanced Learners.* Oxford: Macmillan Publishers Limited.

Sinclair, John 1987. 'The nature of the evidence' in J.Sinclair (ed.) *Looking Up.* Collins: 150-159.

Stubbs, Michael 1996. *Text and Corpus Analysis.* Oxford: Blackwell Publishers.

Stubbs, Michael 2001. 'Texts, Corpora, and Problems of Interpretation'. *Applied Linguistics* 22/2: 149-172.

Turing, A.M. 1950. 'Computing Machinery and Intelligence'. *Mind* LIX.236: 433-460.

Wright, Jon 1999. *Idioms Organiser.* Hove UK: Language Teaching Publications.

Appendix

Entries for **escape** from the New Oxford Dictionary of English (NODE) and the Macmillan English Dictionary for Advanced Learners (MED)

escape ▸ **verb** [no obj.] break free from confinement or control: *two burglars have just escaped from prison*| [as adj. **escaped**] *escaped convicts*
_ [with obj.] elude or get free from (someone): *he drove along the dual carriageway to escape police* _ succeed in avoiding or eluding something dangerous, unpleasant, or undesirable: *the driver **escaped with** a broken knee*| [with obj.] *a baby boy narrowly escaped death.*
_ [with obj.] fail to be noticed or remembered by (someone): *the name escaped him.*| *it may have escaped your notice, but this is not a hotel* _ (of a gas, liquid, or heat) leak from a container _ [with obj.] (of words or sounds) issue involuntarily or inadvertently from (someone or their lips) *a sob escaped her lips.*

New Oxford Dictionary of English (1998)

escape[1] /ɪˈskeɪp/ verb ★★★

1 get away from sth bad
2 avoid sth unpleasant
3 not remember/notice
4 come out by accident
5 go away on holiday
+ PHRASES

1 [I] to get away from a place where you are in danger: *Three people died in the fire, but John escaped through the bedroom window.* ♦ **+from** *His family escaped from Germany and arrived in Britain in 1938.* **1a.** [I/T] to get away from a very unpleasant situation: *people trying to escape poverty* ♦ **+from** *She saw university as a way to escape from her oppressive home life.* **1b.** [I] to get away from a place that you are not supposed to leave such as a prison: *She was shot while trying to escape.* **1c.** [I/T] to get away from an embarrassing or annoying situation: *Maggie started talking to me and I thought I'd never escape.* ♦ **escape sb's clutches** *He was trying to escape the clutches of two amorous young girls.*

2 [I/T] to avoid being killed or seriously injured in an accident or attack: *Two security guards escaped injury in the attack.* ♦ **+with** *Mr Smith escaped with cuts and bruises.* ♦ **escape unhurt/unharmed/unscathed** *Her two-week-old baby escaped unscathed.* ♦ **escape with your life** (=avoid being killed) *He was lucky to escape with his life.* **2a.** [T] to avoid a difficult or unpleasant situation: *The area has escaped the ravages of war.* ♦ *Hughes seems certain to escape punishment.* ♦ **narrowly escape** *Durham narrowly escaped defeat in their first match of the season.* **2b.** [I/T] to avoid thinking about or dealing with an unpleasant situation you are in: **+from** *The cinema allowed people to escape from the depressing realities of their lives.*

3 [T] if something escapes you, you cannot remember it or you do not notice it: *His name escapes me right now.* ♦ *It seems to have escaped him that I was the one who first introduced him to her.* ♦ **escape your attention/notice** *It had not escaped my attention that Joseph was absent.*

4 [I] to come out of a container, usually by accident: *How will we know if there's any gas escaping?* ♦ *About five tonnes of crude oil had escaped into the sea.* **4a.** *literary* to come out of your mouth, although you did not intend it to: *A weary sigh escaped from her lips.*

5 [I] *informal* to go away on holiday: *We're hoping to escape to the Algarve in May.*

there's no escaping the fact that used for saying that something is definitely true or important, even though you may prefer to think that it is not