http://lexikos.journals.ac.za

From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions

Michael Rundell, *Lexicography MasterClass (www.lexmasterclass.com)*
and *Macmillan Dictionary, London, United Kingdom*
(michael.rundell@lexmasterclass.com)

Abstract: Editorial policies and lexicographic conventions have evolved over hundreds of years. They developed at a time when dictionaries were printed books of finite dimensions — as they have been for almost the whole of their history. In many cases, styles which we take for granted as "natural" features of dictionaries are in reality expedients designed to compress maximum information into the limited space available. A simple example is the kind of "recursive" definition found in many English dictionaries where a nominalization (such as *assimilation*) is defined in terms of the related verb ("the act of assimilating or state of being assimilated"), and the user is required to make a second look-up (to the base word). Is this an ideal solution, or was it favoured simply as a less space-intensive alternative to a self-sufficient explanation?

As dictionaries gradually migrate from print to digital media, space constraints disappear. Some problems simply evaporate. To give a trivial example, the need for abbreviations, tildes and the like no longer exists (though a surprising number of dictionaries maintain these conventions even in their digital versions). So the question arises whether we need to revisit, and re-evaluate, the entire range of editorial policies and conventions in the light of changed circumstances. This paper looks at some familiar editorial and presentational conventions, and considers which are no longer appropriate in the digital medium — and what new policies might replace them.

Keywords: DEFINITIONS, EXAMPLE SENTENCES, DIGITAL MEDIA, EXCLUSION CRITERIA, GATEKEEPER, LEXICOGRAPHIC CONVENTIONS, ONLINE DICTIONARY, USER PROFILE

Opsomming: Van druk na digitaal: Implikasies vir woordeboekbeleid en leksikografiese norme. Redigeringsbeleide en leksikografiese norme ontwikkel al oor honderde jare. Dit het ontstaan in die tyd toe 'n woordeboek 'n gedrukte band was met vasgestelde dimensies — soos dit was vir die grootste deel van die geskiedenis van die woordeboek. In baie gevalle is die stylelemente wat as "natuurlike" eienskappe van woordeboeke beskou word, in der waarheid hulpmiddels wat ontwerp is om die maksimum hoeveelheid inligting in 'n beperkte beskikbare ruimte saam te pers. 'n Eenvoudige voorbeeld is 'n rekursiewe definisie wat in 'n aantal Engelse woordeboeke verskyn, waarby 'n nominalisering (soos bv. *assimilasie*) in terme van die verwante werkwoord gedefineer word ("die daad om te assimileer of die toestand van geassimileer wees"), en die gebruiker word genoodsaak om 'n tweede keer (die basiswoord) na te slaan. Is hierdie 'n ideale oplossing of word dit verkies bloot omdat dit minder ruimte in beslag neem as 'n onafhanklike verduideliking?

Soos woordeboeke geleidelik van druk- na digitale medium beweeg, verdwyn hierdie ruim-

tebeperkings. Sekere probleme verdamp eenvoudig. Om 'n nietige voorbeeld te gee, die behoefte aan afkortings, tildes, en dies meer bestaan nie meer nie (alhoewel 'n verbasende aantal woordeboeke hierdie norme selfs in hul digitale weergawes behou). Die vraag ontstaan dus of ons die volledige reeks redigeringsbeleide en norms in die lig van die veranderde omstandighede behoort te beskou en te herevalueer. Hierdie studie neem 'n paar bekende redigeringsbeleide en aanbiedingsnorme in oënskou, en oorweeg dan watter daarvan nie meer toepaslik is in die digitale medium nie en met watter nuwe beleide hulle vervang kan word.

Sleutelwoorde: DEFINISIES, VOORBELDSINNE, DIGITALE MEDIA, UITSLUITINGSKRITERIA, HEKWAGTER, LEKSIKOGRAFIESE KONVENSIES, AANLYN WOORDEBOEK, GEBRUIKERSPROFIEL

1. Setting the scene: from print to digital

This paper revisits a number of familiar and well-established editorial policies and lexicographic conventions. The aim is to discover whether policies which developed during the long period when dictionaries existed only as printed books remain appropriate in the 21st century, when many — if not most — dictionaries are now published in digital media.[1]

In early 2015, in one of the regular updates to its dictionary, Macmillan added entries for 64 chemical elements. This completed the dictionary's coverage of all 118 elements. But it is legitimate to ask why they were not all included in the first place. It is usual practice for dictionaries to cover all members of any clearly-defined set (days of the week, signs of the zodiac, and so on) but in this case it was decided to omit the rarer elements in the interests of including other, more frequent vocabulary items. When dictionaries are published in the form of printed books, editors make decisions like this all the time: a book of finite dimensions sets up a "zero-sum" game, in which the addition of one category of information entails the omission of something else.

The problem can become acute when major new editions are created (typically every four or five years). Newly-emerging words, phrases and meanings need to be added in order to ensure that the dictionary remains current. At this point we have to decide whether to remove an equivalent amount of material in order to accommodate the newcomers (and if so, using what criteria?); whether to increase the size of the book (a popular option, but unsustainable in the long term); or whether to create more space by making typographical adjustments and increasing the amount of text on the page (which may alienate users). Each strategy carries its own risks, which we generally try to minimise through a carefully calibrated combination of all three expedients. For editors of printed dictionaries, the optimal use of limited space is a major preoccupation.

Macmillan's dilemma regarding chemical elements is just one of countless similar decisions forced upon editors working in print media. This is one of the reasons why digital media are so much better adapted as a platform for reference materials of all types (encyclopedias and maps, as well as dictionaries).

Lexicography is going through a turbulent phase and, as dictionaries

gradually migrate from old to new media, "lexicographers … currently live in a sort of interregnum" (Hanks 2015: 87). As always happens when changes are driven by technology, the global picture is uneven. In many parts of the world, paper dictionaries still have a healthy future ahead of them. Furthermore, certain *types* of dictionary — such as those designed for schools, or special-subject dictionaries, or dictionaries of "smaller" languages — may show a preference for print for some time to come. But for three major categories of dictionary — which collectively account for a large chunk of the global dictionary market, and have also had the greatest impact in terms of lexicographic innovation — the long-term decline in sales of printed editions is irreversible and has led publishers to focus increasingly on digital versions. These are: general monolingual dictionaries aimed at adult mother-tongue speakers; bilingual dictionaries for "big" language pairs; and monolingual learner's dictionaries. Progress, for now, is somewhat uneven, but the direction of travel is clear. Nor should this be seen as a cause for nostalgic regret: with unlimited space and digital functions such as multimedia and hyperlinking, new media provide exciting opportunities for innovation and improved coverage, and open up endless possibilities for reference resources which will serve users' needs more effectively than their print-bound predecessors. This paper looks at the implications of this change for the way dictionaries present information and for the type and range of information they include, and asks how well current online dictionaries have responded to the new reality.

2. Background: changes in the publishing model

Though dictionaries in some form pre-date the invention of print (see e.g. Hanks 2010), the dictionaries we are familiar with today largely evolved in the medium of the printed book. For English, this means over 400 years in which editorial policies and lexicographic conventions have developed and become settled. People know what to expect — and what not to expect — of their dictionary: numbered word senses, concise definitions employing familiar (if sometimes incomprehensible) formulae, devices for conveying the sounds of words in written form, and so on. But much of what we take for granted as "natural" features of dictionaries are in reality expedients. They evolved not because they are the best possible way of conveying information to users, but because they satisfy the imperative of shoehorning large amounts of information into a limited space.

Users of Merriam-Webster's dictionaries, for example, will be familiar with their idiosyncratic defining style. This was introduced in the 1950s by chief editor Philip Gove who, according to Kory Stamper (Stamper 2015), was tasked with "saving" 300 pages from the Second International to create the Third. Another source notes that "Every editorial decision Gove made was dictated by space: the need to create as much of it as possible so he could cram new words into the finite boundaries of the printed book. ... Gove claimed he

saved 80 pages in the Third by using fewer commas" (Fatsis 2015). In a typical set of Merriam entries, we learn that *expectant* means "characterized by expectation", and *expectation*, in turn, is defined as "the act or state of expecting". (*Expectancy*, meanwhile, is the subtly different "act, action, or state of expecting".) So a user who genuinely doesn't know the meaning of *expect* (unlikely, but the same "recursive" approach is used for less familiar sets of words too) can only resolve the meaning of *expectant* by making two further look-ups: tiresome for the user, no doubt, but undeniably economical.

The use of telegraphic definitions is not the only space-saving strategy. Until at least the 1990s, the *Concise Oxford Dictionary*, living up to its name, deployed a range of techniques geared to cramming an impressive amount of data into limited space, as this entry for the word *bag* (from the 7th (1982) edition) illustrates:

> **băg**[1] *n.* **1.** receptacle of flexible material with closable opening at top (esp. w. prefixed word showing contents or purpose; DIPLOMATIC *bag*, GAME[1] *bag*, HAND[1]*bag*, KIT[1]*bag*, *mailbag*, *travelling-bag*, VANITY *bag*); (w. such prefix understood) particular kind of this; hence ~FUL. 2 *n.* **2.** contents of bag; MIXED *bag*; amount of game a sportsman has shot or caught (also fig.) **3. ~and baggage**, with all belongings; **~of bones** lean creature**; (whole) ~of tricks** every… [etc]

Figure 1: Partial entry for *bag*, *Concise Oxford Dictionary*, 7th edition, 1982

In the UK, at least, dictionaries had been gradually moving away from these extreme forms of lexicographese even before the move to digital media. But there is always a trade-off, and improved user-friendliness generally meant reduced coverage of the lexicon.

The Macmillan Dictionary, when originally developed for print, had an explicit policy of favouring the most central vocabulary of English. The goal was to provide detailed information (on syntax, collocation, phraseology, register, and so on), backed up by abundant example sentences, for a core set of 7500 high-frequency words. The unavoidable downside of this approach was that words outside this set received more perfunctory treatment, and often lacked examples altogether. (Steps are now being taken to remedy this.) The policy is far from ideal, but it is perfectly defensible in the context of print publishing: adding an example sentence at a word like *parsimonious* could mean that an important pattern at a verb like *instruct* would be left without an example — and for the student who needs to use *instruct* productively, this could be problematic.

Information about morphology is another area where difficult choices have to be made. Among the well-known English monolingual learner's dic-

tionaries, most (in their print editions) provide inflections only for words with irregular morphology: it is assumed, rightly or wrongly, that target users know how the verb *walk* conjugates but may have problems with a verb such as *strive*. The exception is the COBUILD family of dictionaries which have always supplied full inflectional information for every headword, regular or otherwise. Either policy is defensible, but the COBUILD approach carries a space penalty: the three inflected forms shown at *psychoanalyse* take up a full line, and the systematic application of this policy is one of the reasons that COBUILD dictionaries covered a significantly smaller part of the lexicon than their competitors.

If traditional editorial policies and dictionary conventions are — as these cases illustrate — at least partly driven by the space constraints of the printed medium, what happens when those constraints no longer exist?

3. The response so far

The digital revolution has already led to a redefining of what we mean by "dictionary". Contemporary general-purpose monolingual dictionaries now routinely include some or all of the following: a thesaurus, multilingual content, a blog, language-related games or puzzles, "Ask the Editor" features, videos, infographics, and user-generated content of various kinds (Rundell forthcoming). These are supported by almost constant activity on social media. But the focus of this paper is not on the novel features which complement and enhance what is there already, but on the dictionary's central function: describing the meanings and usage of the words in a language. The questions here are how well publishers have adapted to the new medium, how this has affected dictionary macrostructures and microstructures, and what more needs to be done.

In keeping with the uneven way in which innovation is distributed during this transitional period, we currently find several dictionary models co-existing. Broadly, these are: dictionaries published in print form only; those appearing in both print and digital media; and digital-only dictionaries. The second category is probably the most common (at the time of writing), but even digital-only dictionaries are — in most cases — derived from print products. The Macmillan Dictionary is an example of the last type, having started as a printed book in 2002 and moving to a digital-only model in 2013. But the same applies to the so-called "aggregators", online resources such as dictionary.com and thefreedictionary.com. Though apparently "new" products for the digital age, they recycle dictionary data from traditional sources. The smallest category consists of dictionaries conceived and compiled from scratch as digital products. Examples include *Elexiko* ("an online dictionary of contemporary German") published by the *Institut für Deutsche Sprache*; the *Diccionario de Aprendizaje del Español como lengua Extranjera* (DAELE), a Spanish learner's dictionary being developed at Pompeu Fabra University; and the *Algemeen Nederlands Woordenboek* (ANW) being compiled at the *Instituut voor Nederlandse Lexicologie* in Leiden. All are works in progress rather than complete dictionaries,

and all are from non-commercial institutions.

In its most primitive form, a digital dictionary simply makes the text of a printed dictionary available on a website. A notable example is the *Diccionario de la lengua española* (published by the Real Academia española), whose online dictionary is virtually identical to the print product it is derived from. In this entry for *traducción*, no attempt has been made to exploit the possibilities of the new medium: the entry retains the abbreviations, tildes, and recursive definitions ("the act or result of translating") of the original, with no hyperlinks to words referred to in the definitions:

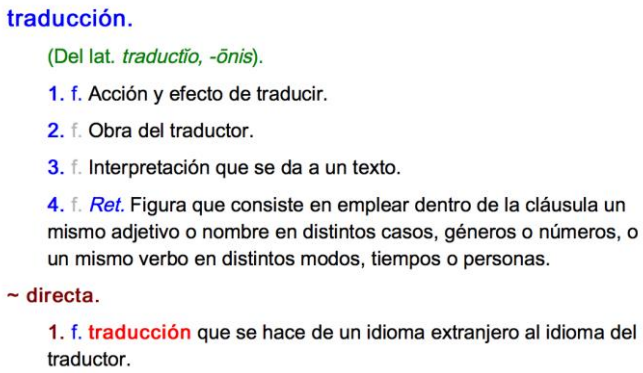

Figure 2: Entry for *traducción*, *Diccionario de la lengua española*

This is an extreme case, but in many online dictionaries, old and new features sit uneasily together. Though never existing in print, *Wordnik* — with its two-column presentation — shows contemporary web-derived example sentences on the right side of the screen, supported, on the left, by definitions from a range of traditional dictionaries. Thus at its entry for *tweet*, we find up-to-the-minute examples of the social-media sense, while the corresponding definitions (derived from an old edition of the *American Heritage Dictionary*) fail to record this recently-coined meaning.

Wiktionary is an especially interesting case. On the face of it, this is a very "modern" dictionary: an entirely web-based resource, its entries created from user-generated content, and with no roots in traditional print lexicography. But things are not quite so simple. Though most entries for subject-specific terminology are newly created, usually by people with specialist knowledge, many of the definitions for more "everyday" vocabulary are simply copied from other dictionaries. Worse, *Wiktionary*'s contributors — rightly concerned about intellectual property issues — tend to borrow material from a safely-out-of-copyright edition of *Webster's Revised Unabridged Dictionar*y published in 1913. Thus many of Wiktionary's entries exhibit long-outdated defining styles and an analysis of word senses which reflects old-fashioned ideas about meaning dating from the

pre-corpus age. As Robert Lew has commented: "It seems that the web community, while enthusiastically embracing the novelty of online collaboration, propagates the traditional model of lexicographic description" (Lew 2014: 17).

In the best "hybrid" dictionaries (where a dictionary created for print publication is also available online), there are still remnants of older ways of doing things. But conscientious efforts are being made to adapt to the new medium. As well as making obvious changes (spelling out abbreviated forms and grammar codes, more "open" design where different information types start on a new line and often in a new colour, and so on), dictionary-makers are rethinking the role of alphabetical order. In a traditional macrostructure, alpha order is the mechanism through which users find what they are looking for. It is so fundamental to the way print dictionaries are organised that early digital dictionaries clung on to this model: they continued to display dictionary entries in alphabetically ordered lists, seemingly reluctant to recognise the irrelevance of this approach in an online resource. But alphabetical order is an arbitrary system which brings together completely unrelated words in sequences like:

redneck, redness, redo, redolent, redoubtable

After some delay, this model is giving way to one more suited to the new medium. The most usual method now is that a search for a specific word brings up the entry for that word and that word only, typically with links to "related words" (as opposed to alphabetically-similar words) shown in a sidebar, as in this entry for *area* from the online version of the *Oxford Advanced Learner's Dictionary*:

Figure 3: Entry for *area*, http://www.oxfordlearnersdictionaries.com

There are residual issues with cross-referencing policy: this entry for *area* still ends with a "see also" list inherited from the print edition, even though the "Other Results" column makes this redundant. But these are teething troubles. A further development relates to phrasal verbs and idioms. These would traditionally be "nested" under main entries, so that *set off*, *set up*, and *set someone's mind at rest* could all be found at the end of the entry for *set*. A newer model — now adopted in many English dictionaries — is to make these items standalone entries. This makes sense if we see phrasal verbs and idioms as distinct lexemes (and many of the former and some of the latter have more than one sense). Why should a user who wants to understand the expressions *put up with* or *set the cat among the pigeons* be obliged to scroll through a long entry before eventually locating their search item somewhere near the bottom?

Some interesting alternatives to conventional macrostructure can be found in dictionaries with no print legacy. This entry from DAELE gives a flavour:

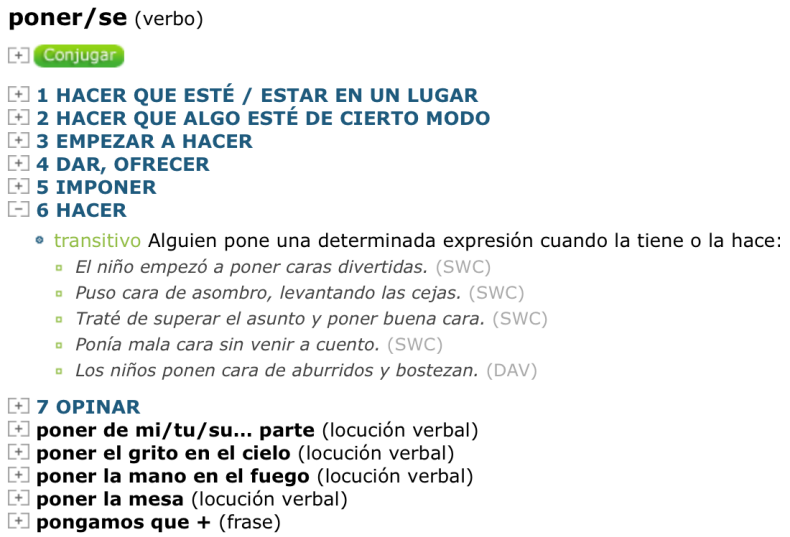

Figure 4: Entry for poner/se, DAELE

Here the "Conjugar" button gives users the option of seeing morphological information, while each of the example sentences comes with information showing the corpus it derives from. But the most notable feature is that each main part of the entry can be opened up or collapsed using the + and - buttons. Starting from a bare menu giving signposts to each sense or usage, the user can pick a specific meaning to see a fuller definition supported by several corpus examples. Some of these features can also be seen in this entry for *Beratung* (counselling or guidance) in *Elexiko*:

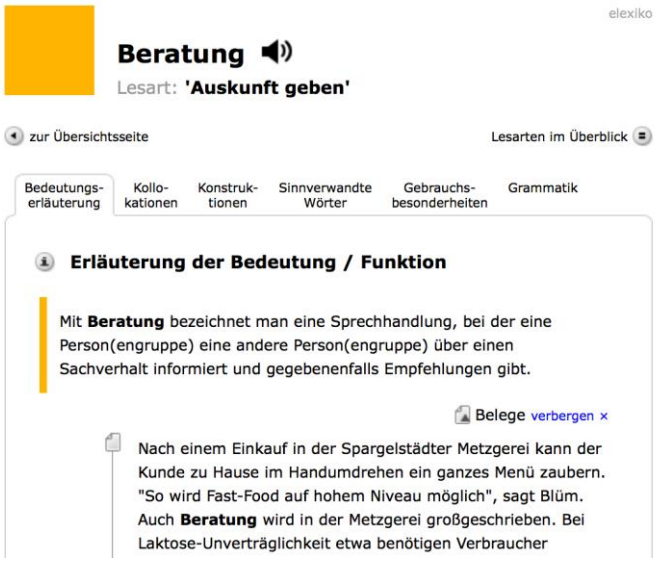

Figure 5: Entry for *Beratung*, *Elexiko*

As in DAELE, users can show or hide examples. And as in DAELE, we find a less traditional approach to defining — in this case a full-sentence "explanation of the meaning or function" of *Beratung*. But what is most interesting here is the use of tabs: these provide access to other categories of information (about collocation, syntactic behaviour, related words, and so on) but at the same time they give users the option of ignoring any information type which they are not (currently) interested in.

Before we conclude this section on current and emerging practice, a few observations are in order. Firstly, there are as yet no "standard" models for the macrostructure of a digital dictionary. What we are seeing at present is a great deal of trial-and-error, as publishers experiment with different approaches. One encouraging sign is the use of generic conventions which can now be assumed to be familiar to anyone using the Web. The + icon in DAELE and the tabs used in *Elexiko* are devices used in many (non-dictionary) websites for managing information, and for suppressing or making available different information-types. The goal in all cases is to avoid overwhelming the user with data, while at the same time making a large quantity of information easily accessible.

The risk of information overload was a challenge for publishers even before dictionaries migrated from print to digital, as the corpus revolution allowed us to provide more complete descriptions of a word's contextual features. To a degree, digital media supply the tools to meet these challenges (even if publishers are still trying to work out the most effective solutions). Part of the problem relates to what Robert Lew has called "presentation space". As Lew

points out, "*storage space* in electronic dictionaries is relatively unlimited", but *presentation space*, which "refers to how much can be presented (displayed, visualized) at a given time to the dictionary user", is self-evidently not (Lew forthcoming). Regardless of how much information the dictionary database contains, the amount that can be presented effectively on a single screen is limited. And with the growing trend for consulting dictionaries on mobile devices, the problem becomes more acute. But dictionary-makers also recognise that users consult dictionaries for different purposes in different situations — broadly, in receptive or productive modes, but with various subdivisions of these types. All these factors intersect, and the challenge for publishers is to design macrostructures which minimise the problem of "too much information", and take advantage of search techniques already familiar to web-savvy users, while facilitating access to different layers of information which will meet different situations of use.

We are, then, in a transitional phase. The challenges and opportunities created by the move to digital media are fairly well understood, and most publishers have grasped the point that the removal of limits on "storage space" is not a licence to abandon the traditional virtues of conciseness. As John Simpson has observed, "if editors were to allow the extent of individual entries to range out of proportion to utility this would result in making the user's task of interpreting an entry much more difficult" (Simpson 2014: 21). There is plenty of experimentation, but as yet little consensus on the way forward.

This is therefore a good moment for a fundamental re-appraisal of policies and conventions which have become so familiar that we may mistake them for being an essential part of any dictionary's DNA. What is needed now is "models for e-dictionaries that focus on critical areas like the data to be included … the structures to present and accommodate the data, the functions of these dictionaries and the way they should respond to the needs of their target users" (Gouws 2014: 157). In the sections that follow, we will look at three specific areas where traditional policies may need rethinking: inclusion criteria, definitions, and example sentences.

Some specific issues: (1) inclusion policies

One of the first questions any dictionary publisher has to consider is "which words get into the dictionary". The theoretical background to this is the observation that the lexicon is an unbounded set. As Hanks points out, "the lexicon is dynamic: new words are being added all the time" (Hanks 2013: 29). When even the mighty OED does not claim to include every English word, it follows that all dictionary publishers need to have robust criteria governing decisions about what to include. But do these criteria need rethinking for the digital age?

Inclusion criteria typically take account of the corpus evidence for a word's frequency, currency, and dispersion across text-types and regions. Some of these criteria are already being modified for digital dictionaries. For exam-

ple, the Oxford Dictionaries site addresses the question of longevity, acknowledging that traditional, stricter criteria regarding how long a word had been current may no longer be appropriate: "It used to be the case that a new term had to be used over a period of two or three years before we could consider adding it to a print dictionary. In today's digital age, the situation has changed" (http://www.oxforddictionaries.com/words/how-do-new-words-enter-oxford-dictionaries).

But two key factors are the "user-profile" of a particular dictionary, and the availability of space. Space limitations require dictionary-makers to be selective about what they include (this has contributed to the public perception of dictionaries as "gatekeepers", only admitting to "the dictionary" words of which they approve), and a good user-profile is the most reliable way of ensuring that the resulting headword list is fit for purpose. A user profile "seeks to characterize the typical user of the dictionary, and the uses to which the dictionary is likely to be put" (Atkins and Rundell 2008: 28). A clear idea of the target user's receptive and productive needs, pre-existing knowledge, language proficiency, and reference skills, is an indispensable aid to inclusion decisions when space is limited.

But neither factor has the same weight when the dictionary is online. It is obvious that unlimited space means inclusion policies can be relaxed, but in an online setting it also becomes much harder to predict who the user will be. In the case of the familiar English monolingual learner's dictionaries, well over 50% of people consulting the site have arrived there through what is known as "organic search": they have submitted to their search engine a search-string (such as "definition of X") which does *not* specify a particular dictionary, then clicked on one of the links in the output. So-called "direct search", where the searcher specifies a particular source (such as Oxford or Macmillan) accounts for a smaller segment of total traffic to most dictionary sites. Consequently, the potential user group is harder to pin down, and this makes it more difficult to feel confident about inclusion decisions.

Samuel Johnson noted rather gloomily that "they that take a dictionary into their hands have been accustomed to expect from it a solution of almost every difficulty" (Johnson 1747: 6). In the digital age, users' expectations are higher than ever. The former "gatekeeper" notion is giving way to a situation where dictionary users (especially younger users) no longer consider that a word is somehow invalid if it is not in "the dictionary". They are more likely to think that if a given dictionary doesn't include a word which they have heard, then the fault lies with the dictionary rather than with the word — and they will simply try a different source. So when there are no space constraints, it may make sense to turn the question around and — rather than asking "does this word pass my inclusion tests?" — we should ask instead "are there good reasons for *not* including this word"? Some traditional principles still apply: candidate words have to be "real" — not invented, or used by only a small group (co-workers, family, or the like) — and they must be supported by independent evidence.

With a general approach based on "exclusion criteria" as our starting point, we need to look at some specific categories. The most difficult of these is so-called "named entities" — broadly speaking, names of people, places, institutions, companies, and so on. Should dictionaries include them? Traditionally, most dictionaries do not include encyclopedic information, but the boundary between encyclopedic and lexical has never been clear-cut, and there is a long list of exceptions. For example, dictionaries generally include the names of institutions which have well-established metonymic uses: thus t*he Kremlin, the Pentagon*, and *Buckingham Palace* will usually feature in any headword list, because corpus data frequently includes sentences like these:

> ***The Kremlin*** *wants the presidential term extended from four to seven years*
> *The Bush White House and* ***the Pentagon*** *seem not to have planned for such contingencies.*
> *How can we be sure that* ***Buckingham Palace*** *has behaved properly in this case?*

The same applies to places real and imaginary which have extended meanings, such as *Mecca* (*At its tip lies Sharm-el-Sheikh, a* ***Mecca*** *for divers and sun-worshippers*) and *Shangri-la* (*a weekend in New York's gay* ***Shangri-la***). In cases like these, dictionaries typically define only the extended uses. Similarly, names like *Google* and *Facebook* only enter most dictionaries as verbs, with no definition for the proper nouns they derive from. There is a host of other quasi-encyclopedic information in dictionaries (such as names of religions, or of trademarked products such as *Band-Aid* and *Memory Stick* which are often used generically). Making these lexical/encyclopedic distinctions is difficult enough for lexicographers, but to the average user they will look arbitrary. But there is clearly a thin-end-of-the-wedge aspect to this. If we decide, for example, to provide definitions for countries (as well as for languages and nationalities, as is the usual convention currently), then why not also for cities, and what is the cut-off point here? And if countries and cities, why not people too — and if so, which ones? The whole issue of which named entities a dictionary should include needs to be revisited in the light of changed circumstances (and changed expectations among users) — though in resources like Babelnet, the lines between encyclopedic and lexical data are already breaking down.

A number of other categories need to be considered. At the end of 2014, the American Dialect Society named as its Word of the Year (WOTY) the social media hashtag #blacklivesmatter. This is a new departure. Dictionary publishers and others routinely nominate Words of the Year, and up to now they have been recognisable as words and have found their way into dictionaries: Oxford's WOTY for 2013, for example, was the now ubiquitous *selfie*. But the American Dialect Society is not alone in extending the scope of what counts as a word. In a readers' poll hosted by dictionary publisher Collins in 2014, the hashtag #nomakeupselfie was a popular choice as the word people most wanted to see in the Collins dictionary, attracting enough votes to come a creditable fourth. Do hashtags belong in dictionaries? Almost certainly not — most are trans-

parent in meaning, and few of them last more than a few weeks — but the question needs to be addressed.

So too does the issue of lexical creativity. This is a pervasive feature of language in use, and lexicographers routinely face inclusion decisions when confronted by examples of it. A newspaper article by the author Margaret Atwood, discussing the concept of freedom in the age of the Internet, provides two interesting examples:

(1) *We human beings have been exploring the border between freedom and* ***unfree-dom*** *for a very long time.*
(2) *Minus our freedom, we may find ourselves no safer; indeed we may be* ***double-plus unfree****, having handed the keys to those who promised to be our defenders.* (Atwood 2015)

Unfreedom is a legitimate formation, but corpus data shows that it is extremely rare (with a hit rate of less than 0.02 per million words) and, being paired here with *freedom*, its meaning is completely clear. Understanding *double-plus unfree* requires a little more background knowledge: the *double-plus* prefix is a feature of Newspeak, the fictional language used by the government of Oceania in Orwell's *Nineteen Eighty-Four*, where it functions as an intensifier. Atwood's choice of words is interesting in the context of her argument, but is any of this lexicographically relevant? The key, as Hanks has argued over many years, is to distinguish between "norms" and "exploitations" (e.g. Hanks 2013: 10-15). In some cases, exploitations can become norms, as one individual's creative coinage gets picked up by others and settles into the language. But such instances are hugely outnumbered by one-off examples of creativity which barely register in corpus data — and which have no place in a dictionary. (See also the discussion here: http://www.macmillandictionaryblog.com/what-goes-in-the-dictionary-when-the-dictionary-is-online.)

The cases discussed above all contribute to fleshing out what we mean by exclusion criteria, and we can now attempt a work-in-progress summary of what these might include:

— user-profile: although (as noted above) this is harder to pin down when a majority of users arrive directly from a search engine, it remains relevant. A general-purpose dictionary is a different animal from a more specialised resource (such as a dictionary of engineering or economics, or a comprehensive historical dictionary), so some filtering is still needed (see also below on technical terms)

— named entities: some broadening of what is acceptable for inclusion seems reasonable, but the question needs more discussion so that robust criteria can emerge

— hashtags: this looks an unlikely category. There may well be a case for a (separate) online resource which lists and explains the most commonly

used hashtags, but there is no real case for these to be included in a general-purpose dictionary

— exploitations: as noted, the traditional position on excluding these is well-founded, and evidence of frequency and dispersion will usually resolve difficult cases

— anything ephemeral: it is always difficult to make reliable predictions about a word's longevity. When dictionaries existed only in print, the problem was less acute. With new editions appearing only every four or five years, editors could often track a word's currency over a longer period. Editors of online dictionaries do not have this luxury, but in principle it is not the function of dictionaries (even those specifically devoted to neologisms) to record the many coinages which appear and disappear in a short space of time

— anything parochial: this is a vague category and not easily defined. But (unlike specialised dialect dictionaries) most dictionaries do not include usages whose range is very limited (whether geographically or socially).

— anything highly technical: the range of specialist "sublanguages" is vast, and few dictionaries (as opposed to specialised glossaries) even scratch the surface in recording their vocabulary. Terminology of the type found in scientific journals like *Nature*, where subject-specialists are addressing other subject-specialists, was rarely included in print dictionaries, and there is no reason for this principle to be relaxed for general-purpose dictionaries in a digital environment.

Individual items can sometimes move unexpectedly into the mainstream. Seemingly ephemeral coinages or parochial usages will in some cases confound our expectations and become part of general vocabulary. Similarly, events in the real world may propel a specialist word towards wider currency: following the global financial crisis of 2008, numerous longstanding technical terms from that sector (*credit default swap, quantitative easing*, and *LIBOR*, among many others) suddenly became part of general discourse — and so merited inclusion in general-purpose dictionaries. But none of this invalidates the broad principles.

Some specific issues: (2) definitions

In section 2 we looked briefly at some of the characteristics of traditional methods of defining. We saw how a focus on economy can lead to definitions which achieve conciseness (and aspire to precision) through the use of standard formulae ("the act of X-ing", "characterised by Y", and so on) and through a "recursive" strategy, where (for example) the entries for *expectant* and *expectancy* feature definitions which are not self-sufficient but depend on the definition at *expect*. And somewhat disappointingly, the user-generated definitions in Wik-

tionary often perpetuate these styles (as does the Spanish dictionary of the Real Academia). In all cases, the goal of saving space is achieved, but the costs are loaded onto the user, who has to learn these conventions in order to fully understand what the dictionary is saying.

In the last 30 years, publishers — especially in the UK — have addressed this issue by developing more open defining styles which approximate to "normal" prose. Even at Merriam-Webster, the digital medium has brought changes in the way words are defined. Its online dictionary (effectively the Merriam-Webster *Collegiate*) provides two layers of definition, as the entry below illustrates: the traditional style — as enjoined by editor Philip Gove — is still there, lower down the entry. But the first thing we see is two new explanations of *expectant* — and these (unlike the so-called FULL DEFINITIONS which follow) require no familiarity with lexicographic conventions and can be fully understood without the need to consult other entries.

Figure 6: Entry for *expectant*, http://www.merriam-webster.com

A related issue is the question of defining vocabularies. Definitions in a learner's dictionary have to be accessible to users with relatively low language proficiency. Most English learner's dictionaries address this issue by identifying a small list (typically of 2000–3000 words) of high-frequency words, and using these, and only these, when writing definitions — the contention being that even quite low-level users will successfully decode any definition. User-research broadly supports this claim, and the use of a defining vocabulary (DV) has been a salient feature of publishers' marketing since these lists were first introduced (see Atkins and Rundell 2008: 449-450). But, for lexicographers and users alike, defining vocabularies are not without their problems, and the cost of clarity can sometimes be a loss of precision. Does the digital medium offer opportunities for improvement? In the digital editions of most (if not all) of the British learner's dictionaries, every word in an entry is hyperlinked. So if a user

is unsure about any word in a definition, they can rapidly find the entry for *that* word. But it is never ideal to have to look from one entry to another in order to get the full picture. A more promising approach may be to create somewhat larger DVs with two or three bands based on frequency. In this model, any word could (and ideally, should) be defined using words from Band 1. But lexicographers would have the option, when necessary, of using words from a higher band — for example, when defining a technical term of low frequency.

As far as the *content* of definitions is concerned (as opposed to the way they are worded), online dictionaries — unconstrained by the need for economy — need to "find the balance between telling the fullest story and deciding what's useful to or necessary for the average reader" (Fatsis 2015). Much has been written about the inadequacies — for many areas of the lexicon — of "classical" approaches to defining, with their insistence on "substitutability" and a model based on *genus* and *differentiae* (e.g. Atkins and Rundell 2008: 416-17). The problem is not merely a practical one — does the definition enable the user to grasp what a word means? — but a theoretical one too. As Hanks has observed "The very word *definition* implies identifying boundaries" (Hanks 2013: 85), and this reflects a traditional view of word meaning which is now being challenged. The assumption that meanings are fixed entities, which "can be attributed to the word in isolation, rather than in context" (Hanks 2015: 87) and which can be described in terms of "necessary and sufficient conditions", is undermined by research in lexical semantics and prototype theory, backed up by the findings of corpus linguistics.

This has led to new approaches to defining (or better, explaining) word meanings. A common thread is a greater focus on context and co-text, and some of these experiments pre-date the digital era.

A notable early example is the "full-sentence definitions" (FSDs) pioneered by the first COBUILD dictionary in 1987, and subsequently adopted (though not systematically) by many other pedagogical dictionaries. This format allows us to include in the definition itself significant (and helpful) information about the definiendum's colligational and collocational preferences, and sometimes also its illocutionary features. I have argued elsewhere (Rundell 2006) that FSDs are not well-adapted to explaining *every* category of word in the lexicon, but they work well in many cases, and they represent an important addition to the definer's repertoire.

On a smaller scale, the Macmillan Dictionary introduced a model for conveying connotative (or pragmatic) information by means of a second sentence. Here a conventional definition explains a word's denotative meaning, then a second sentence adds information about the attitude or motivation of a speaker who chooses to use this word. For example:

> **bureaucrat** someone who is employed to help run an office or government department. This word can suggest that you do not like people like this because you think they have too much power and care too much about rules and systems

Similar examples can be found at the Macmillan entries for *blue-eyed boy*, *nerd*, *just good friends*, *bourgeois*, and many others. Although this style was used in the first (paper) edition of the dictionary (2002), it is clearly well-adapted to a digital structure in which different information types are made available to the user when needed.

In the online *Elexiko* dictionary, traditional definitions are replaced by a statement providing an "explanation of the meaning or function" of the headword. In the case of *Beratung* (see section 3 above), this explanation tells us that it is a speech act, and describes a kind of "frame" in which one person provides another with information about an issue and, where appropriate, makes recommendations.

Yet another innovation can be found in the *Algemeen Nederlands Woordenboek* (ANW) — like *Elexiko*, a genuinely "from scratch" digital resource. Here, conventional definitions are supplemented by "semagrams". A semagram is "the representation of knowledge associated with a word" (Schoonheim and Tempelaars 2010: 721). Thus at the entry for *koe* (cow) we are told about the sound cows make, and we learn that (among other things) they provide milk and meat, have to be milked daily, have udders and four stomachs, and are thought of as being friendly but lazy. All of which "leads to a much richer semantic description, in which the implicit knowledge of the definitions has been made explicit" (ibid.).

A radically different approach is found in Hanks' *Pattern Dictionary of English Verbs* (PDEV), where conventional word senses are replaced by patterns. Here a syntactic pattern (such as *allow someone to do something*) is described in terms of the semantic types (such as Human, Institution, Eventuality) which instantiate the pattern. This description is supported not by anything we would recognise as a definition, but by an "implicature" which maps a meaning onto the specific pattern and its participants. As the site explains, "No attempt is made … to identify the meaning of a verb or noun directly, as a word in isolation. Instead, meanings are associated with prototypical sentence contexts" (Hanks, PDEV).

As all these instances show, dictionary-makers are beginning to explore the possibilities of the new medium. For the time being, there is not much convergence around any new standards, but there are encouraging signs. At one end of the scale, many of the aggregators reproduce material from other sources, with *Wordnik*, for example, featuring definitions from (among others) an ageing edition of the *American Heritage Dictionary*, *Wiktionary* (many of whose definitions come from much older dictionaries), and the truly ancient *Century Dictionary and Cyclopedia*, whose sole definition of *computer* is "One who computes; a reckoner; a calculator". This does not look like a constructive way of exploiting the availability of limitless space. But some of the innovations described here look a great deal more promising.

Some specific issues: (3) example sentences

Dictionary users appreciate example sentences. They help to elucidate meanings, they illustrate contextual preferences, and (especially useful in pedagogical dictionaries) they provide models for language production (e.g. Atkins and Rundell 2008: 452-455). Well before dictionaries went online, the older model of invented, often truncated examples was giving way to the use of authentic examples in the form of complete sentences taken from a corpus. From the 1990s, the provision of *additional* examples became a common feature of dictionaries published on optical disks (CD-ROM and DVD-ROM). Typically these would be taken from a corpus, but in most cases there was little or no filtering (for quality, appropriacy etc) and — critically — examples for polysemous words were not mapped to specific senses.

Now, without the space constraints imposed by the printed medium, publishers of online dictionaries are experimenting with new ways of providing larger numbers of examples. One approach is to give users direct access to the corpora that underpin the dictionary. The *Digitale Wörterbuch der deutschen Sprache* (DWDS), for example, allows users to see concordances in several different corpora, as in this entry for the lemma *Hausarrest*:

Figure 7: Concordances for *Hausarrest*, *Digitale Wörterbuch der deutschen Sprache*

Several other European dictionaries (including the Dutch ANW and the Danish *Den Danske Ordbog*) have a similar concordance feature. This is likely to be a useful resource for linguists and other researchers, but whether non-specialist users want this kind of data (or know what to do with it) is a question which needs to be investigated through user-research.

A different model, which may be better-adapted to the needs and skills of the general user, can be found in *Oxford Dictionaries Online* (ODO). Here, the user will, by default, find one or two examples at most words or senses, but now has the option of clicking on a "MORE EXAMPLE SENTENCES" link to bring up (typically) three further corpus-derived examples. What is especially impressive in the way this is implemented in ODO is that, when the word in question is polysemous, the link appears at individual senses and (as this partial entry for *party* shows), the extra examples are mapped to the meaning which they instantiate:

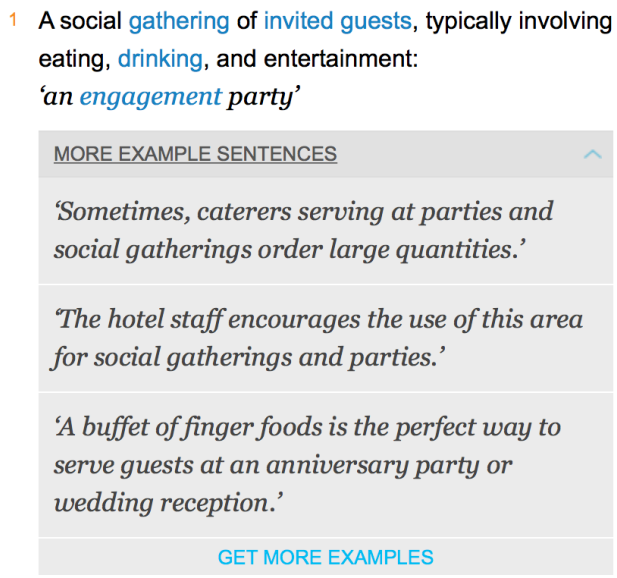

Figure 8: Extra examples feature (*party*) in oxforddictionaries.com

As with the other features, the picture is uneven and not every venture in this area has been entirely successful. As noted earlier (section 3), *Wordnik*'s two-column display (with definitions on one side, and web-sourced examples on the other) runs into problems when the headword has acquired newer senses. In the entry for *toxic*, for instance, almost all of the ten example sentences illustrate more recent uses (toxic assets, toxic relationships etc), but all the definitions on the left side, though taken from five different sources, fail to account for these meanings.

Conclusions

At the time of writing, the online edition of the Spanish Academy's *Diccionario de la lengua española* still defines *diccionario* as "a book" — with no mention in the definition of the medium in which the dictionary appears:

> Libro en el que se recogen y explican de forma ordenada voces de una o más lenguas…

At the other end of the scale, we have a resource like Babelnet. describing itself as "both a multilingual encyclopedic dictionary … and a semantic network which connects concepts and named entities", Babelnet takes full advantage of the digital medium, and provides one model of how "the dictionary" may develop as people experiment with new ways of presenting and linking reference information of various kinds.

This range of responses illustrates how well (and how badly) some dictionary-makers are adapting to the new paradigm. A dictionary is a work-in-progress at the best of times, and as dictionaries steadily migrate to digital media there is a growing flexibility in our ideas about what a dictionary should look like and what information it should contain.

As Gouws has suggested, "The dynamic nature of e-dictionaries enables lexicographers to move away from a static to a dynamic data display that includes the use of a multi-layered structure of dictionary articles" (Gouws 2014: 164). In some of the innovations described here, there is evidence of sensible moves in this direction. One tendency is the increasing use of generic (as opposed to dictionary-specific) conventions for displaying and linking information: hyperlinks, icons for collapsing and expanding a specific category of information, the use of tabs, and so on. A degree of standardisation is emerging in the Web as a whole, and there is a certain "vocabulary" of search strategies which users can now be assumed to be familiar with. So it makes obvious sense for these to be used in dictionary sites, too, since the data on how people arrive at dictionary sites shows that — for many users — the destination is simply an outcome of search, rather than an instance of "looking it up in the dictionary". More generally, dictionary-designers need input from new Web-oriented dictionary-user research and from the field of information science.

But we have also seen that many of these structural innovations are applied to outdated content. Most aggregators recycle entries from dictionaries which pre-date the transformations in lexicography that followed the corpus revolution and the influence of cognitive linguistics. Even a resource as groundbreaking as Babelnet depends, for most of its dictionary content, on Wiktionary — whose definitions of everyday words are in many cases taken from 100-year-old sources. As the scope of the dictionary expands and its structures develop to fully exploit the possibilities of digital media, the lexical data it delivers should also reflect the most up-to-date linguistic thinking about how humans create and understand meanings. This calls for the use of high-

quality corpus-based content, as well as resources such as Hanks' PDEV.

We have looked at three specific areas (inclusion, definitions, and examples) where traditional lexicographic policies are being adjusted to take account of the change in publication medium. This is no more than a first step towards the wholesale re-evaluation of editorial policies and lexicographic conventions which is now needed.

Endnote

1. This paper is based on a talk I gave at the 20th International Conference of Afrilex, held at the University of KwaZulu-Natal, Durban, in July 2015. I am grateful to the Afrilex Board for inviting me as a keynote speaker, and to the conference hosts in Durban for their warm hospitality.

References

Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Atwood, M. 2015. We are Double-plus Unfree. *The Guardian*, September 2015. http://www.theguardian.com/books/2015/sep/18/margaret-atwood-we-are-double-plus-unfree.

Fatsis, S. 2015. The Definition of a Dictionary. *Slate Magazine*. January 2015. http://www.slate.com.

Gouws, R.H. 2014. Article Structures: Moving from Printed to e-Dictionaries. *Lexikos* 24: 155-177.

Hanks, P. 2010. Lexicography, Printing Technology, and the Spread of Renaissance Culture. Dykstra, A. and T. Schoonheim (Eds.). 2010. *Proceedings of the XIV EURALEX International Congress, Leeuwarden, 6–10 July, 2010*: 988-1006. Leeuwarden/Ljouwert: Fryske Akademy.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations.* Cambridge, Mass.: MIT Press.

Hanks, P. 2015. Cognitive Semantics and the Lexicon (Review Article). *International Journal of Lexicography* 28(1): 86-106.

Johnson, S. 1747. *The Plan of an English Dictionary.* Available at https://andromeda.rutgers.edu/~jlynch/Texts/plan.html.

Lew, R. 2014. User-generated Content (UGC) in English Online Dictionaries. *OPAL: Online publizierte Arbeiten zur Linguistik* 2014(4): 8-26. Institut für Deutsche Sprache.

Lew, R. Forthcoming. Space Restrictions in Paper and Electronic Dictionaries and their Implications for the Design of Production Dictionaries. Bański, P. and B. Wójtowicz (Eds.). Forthcoming. *Issues in Modern Lexicography*. München: Lincom Europa.

Rundell, M. 2006. More than One Way to Skin a Cat: Why Full-sentence Definitions have not been Universally Adopted. Corino E., C. Marello and C. Onesti (Eds.). 2006. *Proceedings of the 12th EURALEX International Congress, Torino, Italia, September 6–9, 2006*: 323-337. Alessandria: Edizioni Dell'Orso.

Rundell, M. Forthcoming. Crowdsourcing, Wikis, and User-generated Content, and their Potential Value for Dictionaries. Hanks, P.W. and G.-M. de Schryver (Eds.). Forthcoming. *International Handbook of Modern Lexis and Lexicography*. Berlin: Springer.

Schoonheim, T. and R. Tempelaars. 2010. Dutch Lexicography in Progress: *The Algemeen Nederlands Woordenboek* (ANW). Dykstra, A.and T. Schoonheim (Eds.). 2010. *Proceedings of the XIV EURALEX International Congress, Leeuwarden, 6–10 July, 2010*: 718-725. Leeuwarden/Ljouwert: Fryske Akademy.

Simpson, J. 2014. *What Would Dr Murray Have Made of the OED Online Today?* Slovenščina 2.0, 2 (2): 15-36. Available at http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_03.pdf.

Stamper, K. 2015. *This Wild and Barbarous Jargon, Reduced: Practical Lexicography in an Age of Digital Abundance.* Video of talk given at eLex 2015: Electronic Lexicography in the 21st Century. https://elex.link/elex2015/videos/.

Dictionaries and other online resources

Babelnet: http://babelnet.org.
Digitale Wörterbuch der deutschen Sprache (DWDS): http://www.dwds.de.
Macmillan Dictionary: http: //www.macmillandictionary.com.
Merriam-Webster online: http: //www.merriam-webster.com.
Oxford Dictionaries Online (ODO): http: //www.oxforddictionaries.com.
Oxford Advanced Learner's Dictionary Online: www.oxfordlearnersdictionaries.com.
Pattern Dictionary of English Verbs (PDEV): http: //pdev.org.uk.
Wiktionary: https: //en.wiktionary.org/.
Wordnik: https: //www.wordnik.com.