Crowdsourcing, wikis, and user-generated content, and their potential value for dictionaries

Michael Rundell (March 2015)

Abstract

It is tempting to dismiss crowdsourcing as a largely trivial recent development which has nothing useful to contribute to serious lexicography. This temptation should be resisted. When applied to dictionary-making, the broad term "crowdsourcing" in fact describes a range of distinct methods for creating or gathering linguistic data. A provisional typology is proposed, distinguishing three approaches which are often lumped under the heading "crowdsourcing". These are: user-generated content (UGC), the wiki model, and what is referred to here as "crowdsourcing proper". Each approach is explained, and examples are given of their applications in linguistic and lexicographic projects. The main argument of this paper is that each of these methods - if properly understood and carefully managed - has significant potential for lexicography. The strengths and weaknesses of each model is identified, and suggestions are made for exploiting them in order to facilitate or enhance different operations within the process of developing descriptions of language. Crowdsourcing - in its various forms - should be seen as an opportunity rather than as a threat or diversion

Introduction

An article titled "The Urban Dictionary guide to sex" appeared in the UK's *Guardian* newspaper in 2014 (Benedictus, 2014). It highlighted various recherché terms (naming equally recherché sexual practices), which are described in some detail in the *Urban Dictionary* (hereafter UD), but which have - for some reason - been overlooked by most conventional dictionaries. The piece attracted well over 200 comments from readers, among which was this eye-catching exchange:

> Comment 1: It [UD] has always looked to me like a site for smutty adolescents rather than serious lexicographers.
>
> Comment 2: I don't know what a lexicographer is but it sounds pretty disgusting!

Which puts us lexicographers in our place.

For many (and especially for those who feel uneasy about the whole crowdsourcing enterprise), UD is the prototypical crowdsourced dictionary, and shorthand for the entire genre. In a defence of traditional lexicographic virtues, Jonathon Green takes aim at the *Urban Dictionary* and asks "Do we believe this farrago of misinformation, theorising, one-off terms and a level of 'definition' based on a count of thumbs up and down?" (Green, 2012).

It is easy to see why a serious lexicographer should take exception to this website. The entry for *Republican*, for example, consists of well over 300 "definitions". Some are brief and to the point ("a stupid fascist dumbass"), others run to several paragraphs, and all attract several hundred "likes" and "dislikes". There are numerous ways in which UD differs from a conventional dictionary. Those 300-odd definitions are not describing different senses of the word (which would make *Republican* more polysemous than *take*). Rather, they represent multiple efforts to describe the same meaning, or simply to air prejudices. The Urban Dictionary has nothing approaching a "defining policy" (the noun *domesticity* is defined as "to be overly enamored with all things domestic" - as if it were a verb) and many entries are downright misleading: if you didn't know what a *draughtsman* was, UD's lone definition ("a professional artist of drinking") wouldn't help you much. There appears to be no editorial oversight, no inclusion policy (anything goes), no regularity of style, and no control over extent (the entry for *hipster* is 722 words long, comfortably eclipsing the notoriously rambling definition of *door* in Merriam-Webster's *Third*). You can't even rely on the spelling to be right: *concierge* is defined as "A fancy name for a person that has access to Google and who's job is to …". Indeed, UD's "top definition" of the word *dictionary* appears to concede that it is not a "real" dictionary at all:

> What you're reading right now, but without all the assholes, anti-Americans, dumbass n00bs, atrocious grammar, made up words, slang, gibberish, and other crap.

It would be tempting, on the evidence of the *Urban Dictionary*, to dismiss the whole crowdsourcing enterprise as a superficial Web 2.0 phenomenon. But it is much more than this. In this paper, I will review some of the activities that are commonly lumped under the broad heading "crowdsourcing", and I will attempt to nail down what is currently a vague and ill-defined umbrella term. I will look at examples of crowdsourcing in the fields of linguistics and lexicography, and assess their value for users and producers of dictionaries. And I will discuss the key question of whether crowdsourcing (or some flavour of it) has the potential to make a serious contribution to the development of lexical reference resources.

In the wider world, crowdsourcing is seen by many as an effective new methodology for solving problems, and as "a strategic model to attract an interested, motivated crowd of individuals capable of providing solutions superior in quality and quantity to those that even traditional forms of business can" (Brabham, 2008, p.79) And notwithstanding ob-

vious weaknesses, even the *Urban Dictionary* is by no means without merit. No-one with any sense would use it to find out about "normal" words such as *supercilious, beatify*, or *draughtsman.* But it is one of the best sources around for information about contemporary slang (and significantly, it describes itself quite narrowly as "A veritable cornucopia of streetwise lingo"). And the fact that so many people are ready to contribute their time and opinions suggests that it is more than just "a site for smutty adolescents". The *Urban Dictionary* is not the only, or even the most typical, manifestation of crowdsourcing in the field of lexicography. But it demonstrates the high level of engagement by members of the public in any online discussion relating to language use and language change, and this provides grounds for optimism about the lexicographic potential of crowdsourcing.

(1) Crowdsourcing: a provisional typology

The term "crowdsourcing" is problematic. It is widely used but loosely defined, covering a broad spectrum of working methods. This is a good moment to try to pin the word down, and I will start by identifying three subcategories (not always mutually exclusive) which are typically bundled under this general heading.

The first is "user-generated content", or UGC. A common feature of all crowdsourcing is that it replaces the binary "producer/consumer" model with something more flexible, but this is particularly salient in the case of UGC. In this model, people share their knowledge and expertise (and sometimes their opinions and prejudices), often in an unstructured way with no very clear goal. Secondly, there is the "wiki" approach, which refers to a collaborative method for creating, maintaining, and refining a collection of data, in which self-regulation replaces a central editorial authority. Thirdly, there is crowdsourcing in the narrow sense (hereafter referred to as "crowdsourcing proper"), which denotes a "distributed" model of working, whereby the completion of very large tasks is enabled or facilitated through mass participation.

1.1 User-generated content

UGC is probably the most familiar of these three models. Though pervasive in many areas of the Web, it can also be found in "old media". Television programs, for example, routinely invite viewers to give an opinion or share photos via email or social media. Sports events or popular dramas are accompanied by real-time reactions and running commentary in the Twittersphere. It is a given that audience members are no longer passive consumers but can also be active participants or commentators. The level of activity can be extraordinarily high: an article on Huffington Post will sometimes attract several hundred Comments, and heated debates among readers are a standard feature of online news publishing, and indeed of blogs on any subject.

But there is more to UGC than the sharing of opinions and airing of prejudices. If you want to fix the brakes on your car, learn a card trick, or get the most out of a software package, sites such as WikiHow are full of useful advice. In the past, information like this was typically found in a manual (paper or online) provided by the company which sold you the product. What is different about this new paradigm is that information is often supplied by enthusiasts and expert users - members of the public who contribute their knowledge with no expectation of financial reward. A video tutorial on using the *Macmillan Collocations Dictionary* can be found on YouTube, and - in true UGC fashion - this was produced not by the dictionary's publisher, but by an enthusiastic and well-informed teacher sharing his knowledge.

UGC ranges from attention-seeking trivia to genuinely useful information grounded in real expertise. But regardless of quality, a characteristic feature is its randomness. UGC is less about achieving a specific goal than adding, in an non-directed and ad hoc way, to the sum total of publicly-available knowledge. In the field of lexicography, the *Urban Dictionary* nicely exemplifies this feature of UGC. With no pre-selected "headword list" whose entries need to be filled out, users are free to submit words, phrases, or new definitions at will. Over time, coverage of the lexicon is bound to improve, but there is no guarantee you will find even quite common words; it all depends on the interests and preferences of users. Thus (at the time of writing) UD has no entries for *disport, dissimilar, dissatisfied*, or *dissolute*, and similar gaps can be found throughout the alphabet.

1.2 The wiki model

The wiki model functions quite differently. It invokes the "wisdom of crowds" - the notion that better outcomes can often be achieved by aggregating the views of a large number of people. It is, as Robert Lew explains, an approach which "puts the collective opinion of a group of people above that of a single expert" (Lew, 2014, p.8), and he cites trial by jury as a familiar example of this phenomenon. In this model, there is no presiding "authority" making final decisions about what is right or wrong. Rather, such decisions are the product of an ongoing collaborative process involving a self-regulating community of contributors. Steven Pinker has described language itself as "the original wiki", because it develops from the bottom up and aggregates the contributions of countless speakers in a process which has no end point. Leaving aside this interesting observation, versions of the wiki model have typically been applied to software development and to the creation of informational resources.

The most successful of these is Wikipedia. A well-known report in the journal *Nature* (Giles, 2005) concluded that Wikipedia's scientific articles were broadly comparable, in terms of accuracy, with those of Encyclopedia Britannica. This was followed by a second comparison on similar lines (Casebourne et al., 2012). This more recent analysis - a pilot study for a planned major review - also found that Wikipedia "fared well in comparison with articles from other encyclopaedias", and that articles were "markedly up to date and

well referenced". Commenting on the Giles study, Lew speculates that "Presumably Wikipedia got better rather than worse since that time" (Lew, 2014 p.13). Given the way Wikipedia works, this is a fair assumption. Errors do occur (or are deliberately introduced by politically-motivated contributors), but problems are quickly ironed out by other members of the editing community. Articles are brought up to date with extraordinary speed: thus, within an hour or so of his death (in January 2015), the lengthy article on King Abdullah of Saudi Arabia had been fully updated. And the whole process is remarkably transparent. Wikipedia's guiding principles (its "Five Pillars") are clearly explained, and for every published article, readers can access a complete history of its development, and read all the online discussions among its various editors which underpin the current version of the article.

1.3 "Crowdsourcing proper"

Crowdsourcing, in the narrow definition used here, describes a distributed working method in which a large, centrally-managed task is completed with the help of hundreds, even thousands, of volunteers, each of whom makes a small contribution.

The Andromeda Project is a good example and embodies the main features of this model. The ultimate goal is a detailed mapping of the star clusters in the Andromeda galaxy. Thousands of images generated by the Hubble Telescope are parcelled out among volunteers, and their task is to examine one "small" area of the galaxy in order to identify previously-unknown star clusters. The more usual way of tackling huge tasks like this is to automate the process and get machines to do the legwork. (The way large corpora are part-of-speech-tagged algorithmically rather than manually is a familiar example.) But star clusters elude pattern-recognition software, so the process is not easily automated. Humans, on the other hand, can do the work fairly easily, given a little training and a lot of persistence. The crowdsourced data is fed back to the project, and then further processing is done. As the website explains, "After you help us to find these star clusters, we will use several techniques to determine the age and mass of these objects".

The salient characteristics of crowdsourcing proper are well illustrated here. First, the task is so massive that doing it conventionally would entail unsustainable costs (in time and/or money) - to the point where the task may simply not be feasible. Secondly, the work is not amenable to automation, but can be done by humans who are intelligent and engaged but have no special expertise. And thirdly, the task is well-defined and managed by experts, who typically post-edit or post-process the crowdsourced data. Thus the work of experts and volunteers is complementary, and a certain amount of noise (in the form of suboptimal contributions) can be accommodated. Given these conditions, crowdsourcing can be an effective mechanism for completing a dauntingly large task.

The boundaries between these three subcategories are porous, and there are plenty of "hybrid" cases. TripAdvisor is successful example of UGC, providing an alternative - in the form of hundreds of personal reviews from users - to the promotional websites of ho-

tels and restaurants. But there is a "wisdom of crowds" or wiki element here too: individuals' contributions are aggregated to generate overall scores, and these are continually adjusted as new data from users is added. We also find considerable variation in the degree to which user-generated content is mediated by an editorial authority. In the *Urban Dictionary* there appears to be little editorial oversight. By contrast, *Collins English Dictionary* (on which, more later) retains a strict gatekeeper role, admitting only a fraction of users' submissions to the dictionary itself.

(2) Linguistic and Lexicographic applications

2.1 User-generated content

The random element of UGC is well illustrated in the way traditional dictionary publishers invite contributions from their users. Every entry in ODE (the online Oxford Dictionary of English) ends with the question "What do you find interesting about this word or phrase?", with a box inviting readers' comments. Merriam-Webster has something similar, asking "What made you want to look up [word]? Please tell us where you read or heard it (including the quote, if possible)". Most entries attract no comments (though the feature is fairly new, so that will change), while some have dozens: MW's entry for *voluptuous* has (at the time of writing) no fewer than 42 user comments, such as "I've seen it on a dating site as describing 'body type' and I just wanted to know what it meant. Thanks!". But the chances of things evening out over time (so that every word has its own conversation) seem low, and the poor quality of much of this data (often perfunctory comments like "Cool!" or "OMG") suggests that these features are not primarily motivated by a deep interest in users' linguistic knowledge. Rather, their function is to increase user engagement, because in the online publishing world persuading people to spend more time on your site has positive implications for the revenue which the site produces.

Equally random are the user-generated lists that are a characteristic feature of Wordnik. The dictionary's entry for *element* includes dozens of lists in which this word appears - some sensible and potentially useful (e.g. "Minerals and Mineralogy: List of minerals, elements, group names and geochemistry terms encountered in the science of mineralogy"), others idiosyncratic or inexplicable ("Big Book Gre : An open list of 6703 words by [name]") - but all reflecting the enthusiasm and engagement of their compilers (see also Lew, 2014, p.20).

It is now common for dictionaries to encourage readers to suggest new headwords. For Chambers, the process is basic, with users asked to send their data to an email address. Collins has a more transparent approach, and the thousands of "new" words its readers have suggested are publicly viewable. But its editors apply a strict filter and only a tiny minority of these incomings end up as entries in the main dictionary. Macmillan's strate-

gy is different again. The Macmillan Open Dictionary (MOD), consisting of entries submitted by users, was originally a separate lexicon, but is now integrated with the main Macmillan English Dictionary (MED). As the notes for users indicate, the default policy is to accept submissions provided they meet certain simple conditions. Entries are not "doctored" to conform to the Macmillan defining style, and editors do not intervene provided definitions are correctly spelled and comprehensible, and contain nothing that might cause offence. In the latest version, entries which started life in MOD and were later "promoted" to the MED include information about the original submission: who wrote it and when it was submitted.

All the above are forms of user-generated content. Their lexicographic value is variable, and sometimes the noise-to-signal ratio is too high for them to be of much practical use. They all exhibit the randomness which is an inherent characteristic of UGC, but in the cases described here, data from users can complement the work of professional lexicographers, who retain editorial oversight. Whether this mixed model leads to better outcomes than "pure" UGC (of which the *Urban Dictionary*, compiled entirely by its readers, is the most successful exemplar) is a matter for debate. But we can already see the potential of UGC. How to manage it effectively, for optimal results, will be discussed later.

2.2 The wiki model: Wiktionary

Wiktionary was launched in 2002, and like Wikipedia, it is a collaborative project involving thousands of contributors, with dozens of editions (of varying sizes) for many of the world's languages (Meyer and Gurevych, 2012, pp. 261-2; Creese, 2013, pp. 393-4; Lew, 2014, p.14). Wiktionary's stated goal is to provide enough information to fulfil a decoding (or receptive) function. According to its "Main Page", "We aim to include not only the definition of a word, but also enough information *to really understand* it" [emphasis mine]. There is no mention of Wiktionary's value for *encoding* (or productive) tasks, but this is a challenge which few conventional dictionaries really measure up to (for a fuller discussion, see Atkins and Rundell, 2008, pp. 407-411). Like its sister project, Wiktionary is admirably transparent. It has a Style Guide for contributors, in the form of a series of Help pages. These include a description of the dictionary's inclusion criteria, which are sensible and clearly explained. The section begins with the broad guidance that "A term should be included if it's likely that someone would run across it and want to know what it means", and goes on to flesh out the implications of this general principle. For every entry, there is a Discussion page and a Revision history, enabling users to trace shifts in meaning over time (Creese, 2013, p. 393). And the collaborative wiki model "does not mean…that there is an absence of control", as Hanks notes (Hanks, 2012, p.81). There is an impressive level of editorial oversight, and ill-motivated contributions are quickly weeded out.

There are plenty of positives here. The outstanding success of Wikipedia creates the expectation that Wiktionary could have a similar impact. How far does it fulfil this potential?

One might expect problems with coverage (the words and meanings included), given that Wiktionary - unlike conventional dictionaries - does not start with a pre-existing headword list. There are gaps here and there. Hanks noticed the absence of *rogue elephant* and submitted an entry (Hanks, 2012, p.81); and in an alphabetical stretch between *foundation* and *foundling*, there are no entries for *foundation course, foundation garment,* or *founder member*. But if anything, Wiktionary errs on the side of over-inclusion: in addition to entries for regular inflected forms such as *foundations*, we find words like *foundationalist, foundationalism, foundationally*, and *foundationer*. The first two of these are specialized terms from philosophy - not included in ODE or MW, but reasonably well-attested in corpus data; the second two are well-formed derivational items (hence "possible" words), but there is virtually no evidence to show that they are ever used. Meanwhile, the entry at *founded* (labelled as a verb) is described as follows:

1 past participle of *found*
2 (nonstandard, childish) simple past tense and past participle of *find*
3 To set up; to launch; to institute.
4 Use as a basis for; grounded on.

The first "sense" is valid (but arguably unnecessary - and what about the past tense of *found*?). The second looks like guesswork (wouldn't *finded* be a more likely candidate?). The third is a definition of the verb "to found" (not of the form *founded*). And the fourth is an unsuccessful attempt to account for uses such as "a community founded on Christian principles".

This small sample illustrates the strengths and weaknesses of Wiktionary. Its breadth of coverage is impressive, and particularly strong in the area of terminology (*foundationalist, foundationalism*). On the other hand, the entry for *founded* reveals confusion about the principles of defining. And the inclusion of poorly-attested forms like *foundationally* hints at the absence of any serious corpus basis.

Robert Lew looks in some detail at Wiktionary's description of the verb *to handle* (Lew, 2014, pp.16-17). The first sense is intransitive, defined as "to use the hands" - a doubtful usage which is not recorded elsewhere (even in the OED). The definitions in this entry, Lew observes, "tend to be made up of lists of rather general words, often used in non-contemporary or rare senses", while example sentences "are mostly citations from the Bible or old literary classics, and are all archaic without indicating this fact". Lew compares this entry with its counterpart in LDOCE, which he describes as "a breath of fresh air", with its clear definitions, examples of use that are "contemporary, authentic, and

natural-sounding", and extensive information (missing from Wiktionary) about the combinatorial preferences of *handle*.

Similar problems are evident in Wiktionary's treatment of the noun *condescension*. In a well-known paper, Patrick Hanks showed how the norms associated with this word have changed over time (Hanks, 1998). In current usage it is an unambiguously negative word, defined in ODE as "An attitude of patronizing superiority; disdain". Yet in Wiktionary, the older, neutral meaning comes first, defined as "The act of condescending; voluntary descent from one's rank or dignity in intercourse with an inferior; courtesy toward inferiors". There is so much wrong with this that it is hard to know where to begin. One of the ways in which dictionaries have improved in the last 30 years is in weeding out traditional defining formulae which most users find unintelligible (Atkins & Rundell, 2008, p. 438): good modern dictionaries avoid definitions beginning "the act of …ing", unless there is really no alternative. In any case, the meaning described here is clearly obsolete (but not marked as such), as is much of the language used to define it ("intercourse"!). The entry also records the plural use *condescensions*, a form entirely missing from the British National Corpus, and representing fewer than 1.5% of uses of the noun in a larger corpus. This plural form is grammatically irregular (*condescension* is an uncountable noun), and a good dictionary should inform users of such facts, rather than simply recording oddities or eccentricities that may be found in the data.

As Lew points out, and other commentators have observed, many of these problems can be traced back to the "wholesale incorporation [into Wiktionary] of entries from older out-of-copyright dictionaries" (Lew, 2014, p.14). (The definition criticized here is lifted verbatim from Webster's Revised Unabridged Dictionary of 1913.)

The foregoing discussion highlights three salient features of Wiktionary: its old-fashioned approach to describing word senses; an unhelpful style of defining which has largely disappeared from contemporary dictionaries (apart from the so-called aggregators); and the fact that corpus data has not been consulted at any point in the course of framing the entry. The emphasis is on quantity (high coverage) rather than quality, accuracy, or user-friendliness. As for the corpus revolution, the theoretical insights that have transformed our understanding of how meanings are constructed, the increased focus on users' productive as well as receptive needs, and the move away from "lexicographese" to a more user-focused defining and presentational style - it is as if none of these transformative changes of the last thirty-odd years had ever occurred. Instead, most entries in Wiktionary perpetuate (or revive) outdated lexicographic practices. The irony of this is not lost on Lew, who concludes "It seems that the web community, while enthusiastically embracing the novelty of online collaboration, propagates the traditional model of lexicographic description" (Lew, 2014, p.17).

There is much to recommend in the wiki approach. Its hypertext structure makes it "eminently suitable as a model for the electronic dictionary of the future" (Hanks, 2012, p.82). Its accretive and collaborative nature means that the dictionary is always a work in progress, and this is an appealing template for describing a living language - more appropriate, certainly, than traditional dictionaries with their emphasis on "authority" and implied claims of completeness. It comes as no surprise that Wiktionary is at its best when describing the vocabulary of specialized domains - effectively, when it is closer to the boundary between encyclopedic and lexical knowledge. Meyer and Gurevych's positive take on Wiktionary reflects their special focus on its treatment of terminology. As they observe, "Each contributor has a certain field of expertise. This broad diversity of authors fosters the encoding of a vast amount of domain-specific knowledge" (Meyer & Gurevych, 2012, p.259) - which of course is why Wikipedia is so successful. But while contributors to Wiktionary can be experts on specific subjects, one cannot - without corpus data and the skills to analyze it - be an expert on the words that make up a language's core vocabulary.

2.3 Crowdsourcing proper

The word *crowdsourcing* was coined in 2006, but the practice of enlisting large numbers of volunteers to complete a substantial task is far older. A famous lexicographic example is the "reading programme" established in the UK by the Philological Society in 1857, with the goal of collecting raw data for a new historical English dictionary - what later became the OED. Thousands of readers supplied citations, which the dictionary's editors used as a basis for compiling entries for the new dictionary. In another interesting project - more recent but still "low-tech" - a team of experts from the Summer Institute of Linguistics applied an ingenious crowdsourcing methodology to create a dictionary of Buli, a mainly oral language of rural northern Ghana. Thanks to the work of 30 or so enthusiastic local volunteers, they collected the core vocabulary of Buli (over 10,000 words) in just two weeks, and this was quickly processed into the first-ever Buli dictionary (see Higby, 2013, and the excellent video referred to there).

In both these cases, volunteers were engaged in what they saw as an important cultural project, and they contributed their time enthusiastically. But there are other ways of incentivizing potential contributors. During the 1980s, students at Lancaster University contributed to the transcription of the Lancaster/IBM Spoken English Corpus by completing course assignments: "They were given sections of the recordings it was based on, and were asked to do a phonemic transcription of it as part of their coursework. That transcription was then corrected and included in the corpus" (Tony McEnery, personal communication). More recently, Poesio et al. (2013) have shown the potential of computer games as a mechanism for acquiring large-scale linguistic resources. While manual corpus annotation "still has a place to create resources of very high quality", it is simply not a viable approach when the goal is to process very large corpora. Their game Phrase Detectives was developed to facilitate the annotation of corpora with information on

anaphora resolution - a task which is easy for humans but hard for computers. Games of this type (known as "games with a purpose") are designed to "produce the required resource as a byproduct of the users' playing." (Poesio et al., 2013, p. 3:2).

The most obvious way to motivate contributors is to pay them, and in the linguistics and language engineering communities, Amazon Mechanical Turk, with a robust system for managing payments, is a popular model for gathering research data. Kosem, Gantar and Krek (2013) include a significant element of crowdsourcing, with modest payments attached, in their proposal for a new dictionary of contemporary Slovene. In their project plan, lexicographers retain control of the "analytical and editorial" tasks, but a number of more routine activities, such as evaluating automatically-generated example sentences for collocational naturalness, are devolved to laypeople, who need to have good linguistic awareness but can do the job without any lexicographic training or experience.

As these examples illustrate, crowdsourcing proper comes in a variety of forms. But these diverse approaches share certain common features: there is a large task with a clearly-defined goal; the task is not amenable to automation but can be carried out by intelligent laypeople who do not have specialized skills or training; and the task is managed by experts, who typically do further processing on the data which the task generates.

(3) Motivations: dictionary users and dictionary publishers

The Web and social media have created conditions which have overturned the older, top-down media model, where a small number of providers (whether journalists or lexicographers) delivered expertly-curated content to a large number of consumers. Consumers were for the most part passive: a handful of "Letters to the Editor" of a newspaper (or of a dictionary) represented the limits of user-participation. In the new paradigm, ordinary individuals can make a contribution, and increasingly expect to do so. Robert Lew observes this "urge to be part of an online community, connecting and interacting with others" (Lew, 2014, p.9). Among the vast diversity of these communities, people with a special interest in language are a heterogenous group. They include their share of opinionated ignoramuses, but there are also intelligent and enthusiastic amateurs keen to play their part in improving language resources, and subject-specialists ready to share their expertise by writing scholarly articles for Wikipedia or entries for Wiktionary. In his tirade against the *Urban Dictionary*, Jonathon Green concluded that "If reference is to remain useful then it cannot become amateur hour" (Green, 2012). And when we are faced with some of the nonsense one finds in user-generated content, it is tempting to dismiss the efforts of non-lexicographers as no more than a diversion. But with careful management, amateurs may have a very useful contribution to make.

However, the Web giveth, and the Web taketh away. The emergence of the user-as-contributor coincides with a challenging climate for publishers of commercial dictionaries.

As reference materials migrate from print to digital media, older sources of revenue have started to decline, before an alternative business model has had time to establish itself. Where there is a shortfall in funds for new development, one response is automation. Over the last twenty years or so, many of the tasks involved in creating or maintaining dictionaries have been transferred from humans to computers (e.g. Rundell and Kilgarriff, 2011). If the initial impetus was to relieve editors of tedious, labour-intensive jobs like cross-reference checking, more recent innovations have been driven equally by the need to control editorial costs. GDEX, for instance, a software tool for finding the best candidates for dictionary examples in a set of concordances (Kilgarriff et al., 2008), was initially developed to reduce the cost of a project aimed at creating thousands of new example sentences. But some aspects of lexicography are too difficult for computers (for the time being, at least), while other tasks may be more efficiently accomplished through human post-editing of automatically-extracted data. And in some cases, these human editors do not need specialized lexicographic skills.

This creates an opportunity. We can now envision an approach to dictionary-making in which the work is distributed among three "actors": lexicographers, computers, and volunteer laypeople. Each of the three has its own particular strengths, and the trick is to work out which is the most efficient option for performing a given task. Where volunteers are used, we first need to decide which flavor of crowdsourcing is most likely to deliver the results we need. To a degree, this depends on the type of volunteer we are using, and there are (broadly) two distinct groups: true laypeople who have no special expertise but should be linguistically-aware fluent speakers of the language being described; and subject-specialists who can make an informed contribution to the description of domain-specific vocabulary. Secondly, we must consider what systems of quality control should be put in place to optimize the benefits of user participation (and minimize the amount of "noise" that dictionary editors have to deal with). And finally, we need to think about how best to encourage potential contributors in order to ensure maximum participation.

(4) Horses for courses: how to exploit the potential of user participation

From a user's point of view, the ideal dictionary will do three things: it will include the word the user is searching for; it will provide a description of that word's meaning and usage which is accurate and which reflects what the data tells us about its use in real communication; and it will convey this information in a form that takes account of the user's prior knowledge, so that it is both accessible and comprehensible. How can crowdsourcing contribute to these goals?

The inbuilt quality-control features of a wiki approach make it an effective method for describing specialized lexis (cf. 2.2 above). Volunteers create entries for a term belonging to a field in which they are experts, and members of the wiki community with similar expertise refine or correct the entry if necessary. The wiki model is well-adapted, as we saw earlier, to ensuring high levels of coverage, as subject-specialists create sets of entries relevant to their own field. A simple check against a list of terms in a particular domain

seems to confirm this. A wordlist was extracted from a 60-million-word corpus of environmental science supplied to Macmillan Dictionaries by Lexical Computing Ltd. Coverage of the 297 items in A and B was then compared between Wiktionary and ODE, which generally has excellent coverage of technical vocabulary. Both sources performed well, but Wiktionary had a slight edge, and included four items not present in ODE, including *agroecological*, *aminopyralid, astaxanthin,* and *bioaccumative*. And though the Oxford definitions tended to be clearer and more elegant, those in Wiktionary were generally well written and easy enough to follow.

As discussed earlier, on the other hand, Wiktionary performs poorly in the area of general everyday vocabulary. Sense discrimination is often weak or incoherent, definitions reflect the worst features of older dictionaries, and example sentences are either invented and unconvincing or (if authentic) drawn from non-contemporary sources. None of which is surprising. Disambiguating word senses and crafting definitions are the hardest parts of lexicography, and getting these things right requires lexicographic skills and resources: access to corpus data and the expertise to analyze it, and training in distilling the output of this analysis into well-crafted dictionary entries. Without these advantages, volunteer contributors are, to put it bluntly, out of their depth. They therefore resort to recycling entries from out-of-copyright dictionaries, with predictably disastrous results.

It is reasonable to conclude that a wiki approach has significant potential in the area of domain-specific vocabulary, but is not (in current circumstances) an appropriate mechanism for creating entries for the core vocabulary of a language.

User-generated content has different strengths and weaknesses. In some of its manifestations (like MW's "What made you want to look up [word]?" question), it is primarily a marketing tool aimed at engaging users and creating a sense of community, and there is no harm in this.

More relevant here is the case of dictionaries which encourage submissions of "new" words from their users. CED does this, and seems to attract an average of 30 to 40 suggestions every week (which can be viewed here) - a healthy number. Each is tagged to indicate its status (Candidate, Pending Investigation, or Rejected). Given the perfunctory guidelines for submitters, the site inevitably attracts a great deal of unusable material. Invented portmanteau words feature prominently, and suggestions such as *wrironic* (being wrong and ironic at the same time) and *sangry* (sad and angry) do not seem to plug any obvious lexical gap. Of the most recent 300 submissions at the time of writing, just four are tagged "Candidate" (meaning that the word is likely to be included in CED in its next release), and a handful are rejected outright. The vast majority are still "Pending Investigation" (so vetting these submissions involves a big editorial overhead,) but few look like plausible candidates for inclusion. With such a high noise-to-signal ratio, this seems an inefficient method for identifying neologisms for adding to the dictionary. And although

the submissions page makes clear (rightly) that "all new word additions are subject to an approval process from our editors", a competition in 2014 invited users to vote for their favorites. The word *felfie* (a *selfie* of a farmer, apparently) came second, thus almost bagging a place in CED. Commenting on this in *The Economist*, Robert Lane Greene makes the point that "Lexicographers should be logging the words people actually do use, not the ones they say they like. … it is easy to imagine people voting for a cute coinage they would never actually utter or write" (Greene, 2014; see also Rundell, 2014).

The *Urban Dictionary*, of course, consists entirely of submissions from its users. Its shortcomings are well known and have been discussed above, but it scores well as a record of contemporary colloquial usage - whether new words, senses, or phrases - and the definitions of these items are usually clear enough to convey the meaning. Macmillan's *Open Dictionary*, in place since 2009, has attracted well over 6000 "good" user-generated submissions (a smaller number of incoherent or offensive ones never got as far as being posted). With helpful guidelines for contributors, it receives fewer submissions than CED but a higher percentage of usable material. When a new release of the main dictionary is in preparation, MOD is one of the sources used for identifying novel uses or plugging gaps in the record of already-established vocabulary. A significant proportion of items submitted to MOD - around 30%-40% - eventually get "promoted" to the main dictionary through this process: recent coinages such as *selfie, black swan, CAPTCHA*, *defriend*, and *vape* started life as MOD submissions before being included in the main MED. Many words like this would have found their way into the dictionary anyway, but MOD remains a helpful prompt for editors.

The weakness of UGC, as we saw earlier, is its randomness. The product is dependent on whatever users happen to find interesting or appealing on a given day. While these contributions often have value, they will never provide more than a partial record: UD, for example, has no entry for *surprising* - one of the commonest adjectives in English. But as long as this limitation is taken into account, UGC can have a useful role to play, not only in identifying neologisms, but also in helping us to improve coverage of "long tail" vocabulary, such as domain-specific terms and words from regional Englishes.

Crowdsourcing proper is different again. It is well-established as a means of building resources for lexicography and especially for NLP. It is a useful way of facilitating large-scale tasks that are difficult or impossible to automate. And because its output tends to be raw data which is subsequently moderated or processed by experts, quality control is inherent in the model. The project proposal described in Kosem, Gantar and Krek (2013) nicely demonstrates a point which has come up before: that lexicographic tasks don't necessarily have to be undertaken by lexicographers. Some aspects of the work can be delegated to intelligent laypeople (students, for example), thus conserving the more valuable time of skilled editors.

(5) Conclusions

We have described three varieties of user participation in lexicographic projects: the wiki approach, user-generated content, and what is referred to here as crowdsourcing proper. Each has its own strengths and weaknesses. It is important to be clear about the possible benefits of involving dictionary users in the creation or collection of lexical data, and to identify the specific lexicographic tasks where a given approach can make a useful contribution. Crowdsourcing (in the broad sense) is still something of a novelty in the field of lexicography, and more thought needs to go into maximizing its potential. Partly this means finding more effective ways of incentivizing people to volunteer their time and knowledge, and rewarding them when they do. For example, in a recent modification to Macmillan's *Open Dictionary*, contributors whose words have been promoted to become full MED entries are rewarded with a note at the entry acknowledging their initial submission: their name is thus attached to that entry in perpetuity. Gamification techniques (elements of game playing, such as competition and rewards) may also have a part to play in converting passive users into active contributors.

We also need to set up systems to improve the quality of these contributions. Poesio et al. (2013) describe an imaginative approach to quality control (involving training modules, readily available help pages, and so on), whose goal is to push the process "upstream" (2013, pp.3:19-3:20). Measures like these are designed to improve the quality of the work done by lay contributors, so that the processing load is reduced for the more skilled (and more expensive) editors who post-edit contributors' output. There are lessons here for dictionary publishers. Clear, simple guidelines, and helpful entry forms or entry templates can all help to filter out noise at an early stage in the process. As Judy Pearsall says, "user-generated content, adequately curated and differentiated from core content, can be a viable force for enhancing existing quality content, rather than being seen as merely a marketing strategy or judged as of dubious value" (Pearsall 2013). More work needs to be done in order to achieve an optimal division of labour among the three actors (computers, laypeople, lexicographers) we identified earlier. But there are grounds for optimism about value of crowdsourcing if the process is well managed. There is a resource here, and we would be foolish to ignore its potential.

References

Benedictus, L. (2014). The Urban Dictionary guide to sex: mopeds, porb and awkward arms. The Guardian. March 18, 2014. http://www.theguardian.com/lifeandstyle/2014/mar/18/urban-dictionary-sex-moped-porb-awkward-arm-sexual-slang?CMP=twt_gu

Brabham, Daren C. (2008). Crowdsourcing as a Model for Problem Solving. *The International Journal of Research into New Media Technologies*, 14 (1), 75–90.

Casebourne, I., Davies, C., Fernandes, M., Norman, N. (2012). Assessing the accuracy and quality of Wikipedia entries compared to popular online encyclopaedias: A comparative preliminary study across disciplines in English, Spanish and Arabic. Epic, Brighton, UK. Retrieved from: http://commons.wikimedia.org/wiki/File:EPIC_Oxford_report.pdf

Creese, S. (2013). Exploring the Relationship between Language Change and Dictionary: Compilation in the Age of the Collaborative Dictionary. Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) 2013. *Electronic lexicography in the 21st century: thinking outside the paper.* Proceedings of the eLex 2013 conference, Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/ Eesti Keele Instituut, 392-406.

Giles, J. (2005). Special Report: Internet encyclopaedias go head to head. *Nature* 438 (15 December 2005), 900-901.

Green, J. (2012). Dictionaries are not democratic. *The Guardian*, September 13, 2012. http://www.theguardian.com/books/booksblog/2012/sep/13/dictionaries-democratic-crowdsourcing

Greene, R.L. (2014). Internet Lexicography: A, you're adorkable. *The Economist*, May 22, 2014. http://www.economist.com/blogs/prospero/2014/05/internet-lexicography

Hanks, P. (1998) Enthusiasm and condescension. *Proceedings of the 8th EURALEX International Congress*. Fontenelle, T., Hiligsmann, P., Michiels, A., Moulin, A., Theissen, S. (eds). Liege, Belgium, 151-166.

Hanks, P. (2012). Corpus evidence and electronic lexicography. Granger, S. and Paquot, M. (eds). *Electronic Lexicography*. Oxford: Oxford University Press, 57-82.

Higby, D. (2013). tapping the brain for words. *Macmillan Dictionary Blog*, August 27, 2013. http://www.macmillandictionaryblog.com/tapping-the-brain-for-words

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus, in Bernal, E. and DeCesaris, J. (Eds) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra. 425-433.

Kosem, I., Gantar, P., Krek, S. Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) 2013. *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference, Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, 32-48.

Lew, R. (2014). User-generated content (UGC) in English online dictionaries. OPAL: Online publizierte Arbeiten zur Linguistik. Institut für Deutsche Sprache. 4/2014, 8-26.

Meyer, C. & Gurevych, I. (2012).Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. Granger, S. and Paquot, M. (eds). *Electronic Lexicography.* Oxford: Oxford University Press, 259-292

Pearsall, J. (2013). The future of dictionaries. Kernerman Dictionary News. 21 (July 2013), 2-4.

Poesio, M. (2013): Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, Vol. 3, No. 1, Article 3, Pub. date: April 2013.

Rundell, M. (2014) What goes in the dictionary when the dictionary is online? *Macmillan Dictionary Blog* July 15th 2014 http://www.macmillandictionaryblog.com/what-goes-in-the-dictionary-when-the-dictionary-is-online

Rundell, M. and Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In Meunier F., De Cock S., Gilquin G. and Paquot M. (Eds), A Taste for Corpora. A tribute to Professor Sylviane Granger. Benjamins: 257-281.

Dictionaries

CED Collins English Dictionary www.collinsdictionary.com/dictionary/english

ODE Oxford Dictionaries Online: English www.oxforddictionaries.com

LDOCE Longman Dictionary of Contemporary English http://www.ldoceonline.com

MW Merriam-Webster online http://www.merriam-webster.com

MED Macmillan Dictionary http://www.macmillandictionary.com

MOD Macmillan Open Dictionary http://www.macmillandictionary.com/open-dictionary/

UD Urban Dictionary http://www.urbandictionary.com

Wordnik https://www.wordnik.com