Issues in collaborative and crowdsourced lexicography

Franck Sajous and Amélie Josselin-Leray
CLLE, CNRS & Université de Toulouse 2

This document is the authors' version of the book chapter published in *The Bloomsbury Handbook of Lexicography* (ISBN: 978-1-3501-8170-0) : `http://dx.doi.org/10.5040/9781350181731`

To cite this paper:
Franck Sajous and Amélie Josselin-Leray. (2022). Issues in collaborative and crowdsourced lexicography. In Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography*. London: Bloomsbury Academic, pp. 343–358.

Issues in collaborative and crowdsourced lexicography

Franck Sajous and Amélie Josselin-Leray

1 Introduction

Within a few decades, lexicography has undergone a number of changes, be it from a theoretical, technological or economic point of view. The major ones can be listed as follows: the descriptive revolution (Trap-Jensen, 2018), the computerization of print dictionaries (Nagao et al., 1980; Berg et al., 1988), the contribution of corpus linguistics (Rundell and Stock, 1992) and the NLP tool-assisted data analysis (Rundell and Kilgarriff, 2011), the release of various forms of e-dictionaries and their online publication (Nesi, 2008) and, finally, for some, the end of print versions (Rundell, 2014). While some of these changes result from internal shifts triggered by the private and the academic sectors, some others have been caused by external factors. Among the latter, according to Gao (2012), one can find the rise of several types of free online dictionaries such as 'potpourri of dictionaries' (dictionary aggregators) and 'DIY dictionaries' (e.g. *Wiktionary* or *Urban Dictionary*). With such dictionaries being free, commercial dictionaries have had to adapt their business model (Kilgarriff, 2005). The emergence of these new resources also raises questions about new ways of compiling dictionaries. 'DIY dictionaries', which are described either as 'collaborative' or 'crowdsourced', are evidence of the interest of the crowds for lexical descriptions; they also show that internet users can contribute in various ways to self-organized amateur lexicography projects. Another simultaneous innovation is that other disciplines, such as NLP, have started to resort to microtasking – an implementation of crowdsourcing, also referred to as microworking, that consists in breaking down a complex task into simpler tasks that can be performed by various workers online – for annotation projects. Professional lexicographers who consider resorting to volunteers for dictionary compiling may draw inspiration from such approaches. However, one may wonder how the crowdsourced data production or annotation experiments that took place in NLP, for which little or no prior knowledge is required, can apply to the context of lexicography. Several closely related questions arise: can the crowds be guided within an institutional

framework? Which type of lexicographic project can they be involved in? Which tasks could/should they perform, and when they do, within which participatory schemes? When it comes to tasks requiring greater linguistic competence – or, at least, sensitivity – to what extent can the analysis of 'DIY dictionaries' give an accurate and complete picture of what amateurs are able to produce?

Establishing how relevant it is to resort to the crowds and which implementation is more suitable depends on the very nature of the lexicographic project (which type of dictionary?), on its degree of completion (is it a new dictionary being compiled or an existing dictionary being updated?) and on the resources (corpora and tools) that are available for the language under study. This chapter is based on the analysis of several projects which rely on various schemes involving the crowds, either on an experimental or on a large scale. It aims at describing the ins and outs of such collaboration- or crowdsourcing-based lexicographic projects, focusing in particular on their potential, their challenges and their limitations.

Section 2 tries to identify what the notions of *crowdsourcing* and *collaboration* encompass. Section 3 distinguishes the (supervised) processes that aim at dictionary writing *with* the crowds from the (autonomous) processes that rely on their writing *by* the crowds. After considering the reasons of using the crowds in the lexicographic process, where many tasks can be automated (Section 4), in Sections 5 to 7 we study three types of projects that can benefit from this external help: traditional institutional projects in which volunteers are entrusted with annotation tasks that take place during the data analysis stage, projects based on field linguistics, i.e. the collection of linguistic data where amateurs are considered as informants, and open dictionary projects where users are asked to suggest additions and modifications. Finally, Section 8 addresses the ethical problems that can arise from the different implementations of the work done by the crowds.

2 What is *crowdsourcing* and what is *collaboration*?

The adjectives 'collaborative' and 'crowdsourced' are commonly used – sometimes interchangeably – to refer to projects fed by the crowds, but a distinction needs to be made between the two. A term that was made popular by Howe (2006), *crowdsourcing* originally referred to the outsourcing by companies of tasks to be performed by the crowds, i.e. communities of internet users. It has now become an umbrella term which encompasses several categories of methods that are used in var-

Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography*.
London: Bloomsbury Academic, pp. 343–358

ious fields. It actually takes Estellés-Arolas et al. (2015) a 120-word long description to provide an integrated definition of crowdsourcing, which comes after no less than forty – often divergent – definitions of the concept, all taken from the literature. The list of key ingredients underlying the concept have been summed up by Brabham (2013: 3) as follows: 'an organization that has a task it needs performed, a community (crowd) that is willing to perform the task voluntarily, an online environment that allows the work to take place and the community to interact with the organization, and mutual benefit for the organization and the community.' The author (ibid.: XV) also attempts at giving a more concise definition: '[the] deliberate blend of bottom-up, open, creative process with top-down organizational goals.' Among the various crowdsourcing approaches, *microtasking* is a form of distributed work which consists in breaking down a problem that needs to be solved (or data that need annotating) into a large number of simple tasks which will then be assigned to several *microworkers*; those microworkers will receive a minimum amount of money for performing the tasks which will then be aggregated to produce the final result. This approach is based on the search for redundancy and consensus. The same task is assigned to several microworkers, and the result is considered reliable only if the contributions converge. For some lexicographic projects, the integration of microtasking – which is already common practice in NLP – into the overall dictionary-making workflow has started being seriously considered (Čibej et al., 2015). The collaboration approach relies on interaction between several people (e.g. between contributors only or between contributors and the organizing body) who intend to achieve the same goal (even though different individual objectives might also be involved).

The presence of interaction is what differentiates collaboration from microtasking the most. Two more differences, which are related to each other, can be mentioned: what motivates the internet user to perform a given task and how well he/she is familiar with the aim of the task (i.e. which overall project does the task fit into?). As far as microtasking is concerned, microworkers seldom know what the answers they provide will be used for and their motivation is mostly a financial one. Conversely, in collaborative projects – whether they are dictionaries which are compiled outside an institutional framework, like *Wiktionary*, open dictionaries or instances of 'field lexicography' (see Section 6) – in most cases, contributors are aware of the overall intended purpose, and their only or main motivation might consist in achieving this very objective.

The notions of collaboration and distributed work are by no means new; nor are they mutually exclusive. An early implementation of crowdsourcing is the reading programme of the *Oxford English Dictionary*

Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography*.
London: Bloomsbury Academic, pp. 343–358

(OED),[1] which was launched in 1857 to collect a corpus of quotations. It recruited voluntary and paid readers who would copy contexts of occurrences for a number of words and would then send the slips of paper by post. In a more recent context, the compiling of a dictionary by lexicographers who work from a remote location and who have been assigned a number of entries to write could be considered a form of distributed work which is akin to crowdsourcing. But at the same time, in the cases when the definitions written by a given lexicographer are systematically reviewed by another, the process could be understood as collaboration. What makes the recent approaches based on the contribution of amateur crowds innovative is the number, diversity and range of skills of the individuals that are involved, as well as the ways the various participatory schemes are implemented. Which types of approach and which types of contributors are most suitable for a given task deserves further investigation, together with an analysis of the contributors' motivation. To what extent can the study of dictionaries compiled by the crowds provide some possible answers?

3 Dictionaries written *by* the crowds vs dictionaries written *with* the crowds

This section tries to determine whether dictionaries written *by* the crowds can shed light on the best way to write dictionaries *with* the crowds within an institutional framework. Even though there is an obvious link between dictionaries fed by the crowds and the process of collaborative or crowdsourced writing (since the former are the result of the latter), both dictionaries and process are of interest for two different fields: metalexicography (which focuses on the end result, i.e. the dictionary) and lexicography (which focuses on the process itself). At this point, it seems necessary to make it clear that this chapter focuses mostly on the dictionary-making process (i.e. lexicography). Metalexicographic studies are nevertheless worth mentioning since their analysis and description of dictionaries provide valuable insights into the lexicographic processes. Interpreting and generalizing the findings should be done with caution, though, since dictionaries written by the crowds are a complex object of study, as will be shown below.

Some features of dictionaries such as *Wiktionary* and *Urban Dictionary* have been identified through quantitative and qualitative analy-

[1] This type of parallel is refuted by Brabham (2013:9-10) on the grounds that 'crowdsourcing is not old [...] it is a new phenomenon that relies on the technology of the Internet.' His only argument to justify his viewpoint is the fact that the Internet 'make[s] crowdsourcing qualitatively different from the open problem-solving and collaborative production processes of yesteryear.'

Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography*.
London: Bloomsbury Academic, pp. 343–358

ses. For example, Meyer and Gurevych (2012) quantitatively analysed several editions of *Wiktionary* and compared them to other resources. They chose to use resources available in electronic format (e.g. *WordNet*) in order to automate their comparisons. By characterizing the size of the various headword lists, or the correlation between the number of senses by lexical unit in the various resources, what they actually assess for *Wiktionary* is its capacity to act as a lexical resource for NLP, and not its value as a dictionary for humans – which it is in the first place. Qualitative studies were led by Hanks (2012) and Rundell (2017), who commented on the definitions of the English *Wiktionary* by analysing a limited number of examples. Even if we may think they went through a larger number of definitions than what appears in the papers, we may wonder what the size of a representative sample could be given the size of the headword list of this dictionary. Following an 'old-fashioned approach to describing word senses' (in particular, a large number of derivative definitions), they explain that the definitions under scrutiny are taken from dictionaries which are old enough to be copyright-free – which was also pointed out by Sajous and Hathout (2015) regarding the French *Wiktionary*. In the same way, Sajous et al. (2019) showed that the alternating presence/absence of point of view in the English and the French *Wiktionary* is mostly due to the import of entries from existing dictionaries. In other words, *Wiktionary* entries might not necessarily reflect the lexicographic skills of amateurs, but more specifically the features of dictionaries from the past. Some areas of the lexicon, however, prove particularly useful for analysing the specific contributions of the crowds: (i) neologisms found in the general language, and (ii) recent specialized terms, whose treatment cannot be ascribed to older sources. According to Sajous et al. (2020), the French *Wiktionary* can claim a better coverage of the lexicon of computer science than a commercial, general-purpose dictionary, and the definitions of terms pertaining to that field are more accurate. Another study by Sajous et al. (2018) has shown how swiftly amateurs are likely to detect formal and semantic neology in *Wiktionary* but also in *Urban Dictionary*. The latter, which was originally designed as a slang dictionary and which is known to have become a virtual playground and an escape valve for some – which it actually is – sometimes turns out to be the only lexicographic resource available that includes the type of knowledge which is required to fully understand the meaning of some lexical units from a given field or subculture. There is no denying that the dictionary's policy, which encourages contributors to express their points of view, combined with a form of editing control which does exist but can be deemed inefficient, paves the way for a large number of inside jokes and hate speeches. However, it also generates a large number of metalinguistics remarks targeting occurrences of misuse

Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography*.
London: Bloomsbury Academic, pp. 343–358

of the lexicon. Relevant analyses of some polysemous lexical units, which also include a diachronic description, can also be found. In a nutshell, as stated by Damaso (2005: 59), *Urban Dictionary* is both 'a toy and a weapon', but also 'a tool'. Obviously, not all relevant pieces of information found in *Urban Dictionary* have to be recorded in an institutional dictionary, but they do show that some contributors have real analysis skills, and also bring extra information – in their own way – to more conventional lexicographic descriptions.

Even if routine tasks involved in professional lexicography can be fruitfully performed via microtasking, confining the crowds to those 'menial tasks' might not be the only option. The clear-sightedness and linguistic intuition of contributors can also be put to good use, especially in open dictionaries or in field linguistics projects, as shown below.

4 Do lexicographers need the crowds (when they already have corpora and tools)?

Rundell and Kilgarriff (2011) and Kilgarriff (cf. Chapter 7) give an overview of the tools available for corpus lexicography, of the tasks they can perform automatically and of those that can be partly automated as a support for lexicographers: the compiling, cleaning and annotating (lemmatization and POS-tagging) of corpora; the building of headword lists (word frequency counting, detection of formal neologisms); collocation calculation; lexical profiling; the visualizing and sorting of occurrences (concordancers, choice of good examples); vocabulary tagging (assigning of grammatical tags based on syntactic annotation, of field tags based on the corpus metadata), etc. Since the lexicographer seems to be relieved of the most tedious tasks thanks to automation and is 'only' left with the actual writing of the dictionary entries, one may wonder how relevant resorting to the crowds can be. There are in fact four main arguments in favour of involving the crowds. First, corpus lexicography relies on NLP tools which are based on machine-learning systems that use datasets – which often happen to be crowdsourced – either in the training phase or in the evaluation phase. Second, no matter how much these tools can be improved, they will never be flawless. There is noise in the input data, and noise in the output data. Paradoxically enough, tools have allowed lexicographers to save some time, but, simultaneously, their ever improving processing capacities have also exponentially increased the amount of data to be analysed: it is necessary for the results that are automatically obtained to be manually validated or invalidated. Such a lengthy and tedious process sometimes requires minimal language skills and can be accomplished, under certain condi-

Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography.*
London: Bloomsbury Academic, pp. 343–358

tions, by the crowds. Third, some tasks still cannot be automatically undertaken, as underlined by Rundell and Kilgarriff (2011): 'Automated lexicography is still some way off. In particular, we have not yet reached the point where definition writing and (hardest of all) word sense disambiguation (WSD) are carried out by machines.' Despite the studies that have been carried out since then, their remark still stands today. One may wonder if the crowds, rather than editing the output of the tools, could not simply replace them. Fourth, the automation of tasks by tools is only possible when a given language has digital corpora and tools specifically designed to process them. When there are none, corpus lexicography has to be replaced with another type of project which relies on field linguistics, for which one can appeal to crowds of informants in pioneering ways. Finally, once the dictionary compiling process is over, lexicographers can call upon the crowds for user feedback and updating advice.

The three following sections describe the various stages in which the crowds can be involved, depending on the type of project and its degree of completion.

5 Integrating the crowds into the professional lexicographic process

5.1 Crowds + NLP

Within the context of a monolingual Slovenian dictionary project, Kosem et al. (2013) integrate a crowdsourcing task aimed at identifying false positives among automatically extracted collocations, or bad examples, i.e. examples where the collocations do not appear in the expected syntactic structure. The examples have been randomly drawn from a gold standard that has been designed specifically for the task and are presented to participants who need to assess their reliability. According to the authors, the experiment, which was in the experimental phase at the time, produced highly reliable results (no figures are provided). Following on, Čibej et al. (2015) consider integrating crowdsourcing into the overall workflow of lexicographic projects. They draw up a list of tasks in which the crowds could be involved and list recommendations for the development and implementation of the corresponding microtasks. All the tasks that may be crowdsourced deal with the data analysis phase, while the editorial work remains in the lexicographer's hands. Kosem et al. (2018) take up the task of identifying false collocations which was described above. The new experiment involves 4 participants who annotate 6,590 collocations for 88 sample headwords, through microtasks presented via an in-house interface. The results, which, in this study,

were quantified, show an encouragingly high inter-annotator agreement, but this measurement alone is no guarantee for the quality of the results, as will be shown in Section 8.2. One issue raised by the 2013 and 2018 experiments is what a scaled-up version would be like in terms of participants. In the NLP field, the experiments carried out through microwork platforms exclusively deal with the English language. We are not aware of any large-scale language annotation experiments carried out through microwork for any other language. In the case of Kosem et al.'s (2018) experiment, the authors write that the annotation tasks they propose are 'not very demanding, even for non-linguists,' but their annotators are students in linguistics. Kosem et al. (2013) use non-lexicographers 'with good knowledge of a language.' All these experiments can be relevant as proof of concept, but raise questions about the possibility of broader recruitment. Could this type of task also be performed by naive people? If so, is a more massive recruitment of speakers of Slovene (and more generally, of other languages) conceivable? If not, do the authors have a sufficiently large pool of student linguists?

5.2 Crowds vs NLP

Even today, many data analysis tasks remain difficult to automate using NLP tools. Two of those tasks –definition writing and WSD– are already mentioned by Rundell and Kilgarriff (2011). Two additional ones that seem even harder to undertake are (i) Word Sense Induction (WSI), a preliminary phase which consists in identifying the different meanings of a lexical unit, and (ii) the detection of semantic neology. This section tries to establish whether, for performing such complex tasks, mobilizing the crowds could be an alternative to designing new algorithms. Some unsupervised clustering algorithms (in particular topic-modelling algorithms) tackle the task of WSI by grouping the contexts in which lexical units appear in a given corpus, but lead to some problems. On the one hand, many require to determine a priori the number of clusters associated with each lexical unit (each cluster ultimately corresponds to a given sense). Some papers, such as Lau et al. (2012) propose solutions whose algorithm tries to find out what an appropriate level of granularity would be. On the other hand, it is very difficult to anticipate what the optimal parameterization for this type of algorithm could be, especially since the evaluation procedures are complex, as shown by the SemEval-2010 campaign (Manandhar et al., 2010).

Although it is more common to replace humans by machines, using amateur crowds where algorithms perform poorly can also be considered. Microwork is suitable for simple tasks. For complex tasks, procedures that automatically break them down into simpler subtasks may be

developed. Rumshisky (2011) proposes such a strategy based on microtasking 'intended to imitate the work done by a lexicographer in corpus-based dictionary construction' for WSI and WSD. With this goal in mind, she designs an iterative process that groups together occurrences deemed to have a similar meaning. The process consists in presenting microworkers, for a given word and a target occurrence, with all the other occurrences one after the other. The microworker must determine whether the meaning of the word in context is similar to that of the target occurrence. The occurrences selected by majority vote form a cluster with the target occurrence. Not only does the proposed strategy generate a sense inventory and a sense-annotated corpus, but it also provides metrics based on the inter-rater agreement/disagreement that estimate the coherence of each cluster, the typicality of an occurrence for a given cluster, and the proximity between two clusters.

As mentioned earlier, another task which is considered difficult to automate is the detection of semantic neology. Lau et al. (2012) suggest adapting a WSI algorithm to discover new word meanings, which they apply to the ukWaC corpus (focus corpus) and the BNC (reference corpus). Cook et al. (2013) apply this method to newswire articles taken from the Gigaword corpus and ask an experienced lexicographer to analyse the results. Even if false positives are proposed (and if it can be assumed that proven cases of semantic neology are overlooked), the evaluation shows the relevance of integrating such a system into the lexicographer's toolbox. As far as distributional semantics is concerned, prediction models based on neural embeddings which have recently been used for the detection of semantic neology raise the same issues as count models based on explicit distributional vector spaces, such as those implemented by Gulordava and Baroni (2011): 'they do not account for polysemy, and appear best-suited to identifying changes in predominant sense' (Lau et al., 2012). Words embeddings are commonly used to detect semantic shifts between two synchronic corpora which differ in nature (e.g. different genre/domain). For instance, Fišer and Ljubešić (2018) attempt to differentiate standard and non-standard Slovenian by comparing embeddings learnt from the contemporary Gigafida and Tweeter corpora. The embeddings used to detect semantic neology – diachronic embeddings – are built in the same way as those intended to detect semantic shifts between synchronic corpora. For example, Hamilton et al. (2016) use GoogleBooks N-Grams over the 1800-1999 period, which they divide up into 10-year time periods. Regardless of the very specific nature of the Twitter and GoogleBooks corpora, detecting semantic neology requires to fulfill two opposite needs: (i) reaching a critical volume of data that can be exploited by neural models while (ii) limiting the texts under study to time periods that are sufficiently short (e.g. one or two years)

Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography*.
London: Bloomsbury Academic, pp. 343–358

to detect semantic shifts that are recent enough for lexicographic use. As far as GoogleBooks N-Grams are concerned, it should be noted that they are not released on a regular basis – the last version to be released before 2020 was the 2012 version. As with WSD and WSI, semantic neology detection is an area where the crowds may very well compete with algorithms: in the same way as Rumshisky (2011) adapt a WSI method for the detection of new meanings of lexical units, Rumshisky's (2011) iterative method, which uses crowdsourcing to infer a sense inventory and a disambiguated corpus, could very well be adapted to the task. An estimate of the time and cost involved needs to be made, but compiling a corpus in keeping with this approach seems more feasible than compiling one for the construction of diachronic embeddings.

To sum up, several interesting approaches have been proposed to undertake a number of difficult tasks: WSI, WSD, and semantic neology detection. However, their implementation, whether based on automation or crowdsourcing, is still perfectible and the viability of a large-scale integration into a lexicographic project remains questionable. In the meantime, turning to dictionaries entirely written by the crowds might offer new prospects: in 2012, Lau et al. gave two examples of new meanings 'not included in many dictionaries': 'send a message on Tweeter' for the verb *tweet* and 'style' for the noun *swag*. The new meaning of *tweet* (added to the OED in June 2013) was recorded in *Wiktionary* on Feb. 22, 2009 (as a reminder, Tweeter was launched in 2006). *Swag* (n. 2) was a new entry added to the OED in January 2018 but it first appeared in the *Macmillan Dictionary* in August 2012 thanks to its open crowdsourced dictionary, and in *Wiktionary* on Oct. 8, 2011. Fully collaborative dictionaries and crowdsourced ones tend to include formal neologisms quickly and extensively, but also to record semantic neology (Sajous et al., 2018). Whether they are automatic or crowdsourced, the methods for detecting neologisms could therefore be complemented by careful scrutiny of dictionaries such as *Wiktionary* and, to some extent, *Urban Dictionary*. A hybrid solution may be considered in the future: either by using crowdsourcing before using methods such as those developed by Lau et al. (2012) and Cook et al. (2013) (which brings us back to the above-mentioned 'crowds + NLP' configuration), or by automatically cross-checking data taken from dictionaries written by the crowds with those obtained by automatic corpus processing.

6 Lexicography and field linguistics 2.0

The involvement of the crowds in the compiling or updating of dictionaries as described in the previous section only holds in projects for which digital corpora and tools are available. For the lexical descrip-

tion of languages that have neither (e.g. Swahili or Zapotec languages), data collection must be carried out beforehand, or simultaneously if the dictionary is being published (online) while it is being created. The data collection phase is a field linguistics task that is traditionally performed by linguists/lexicographers 'in person', together with the informants, but that can also benefit from the use of online tools, as illustrated below.

In the Kamusi project, whose objective is the production of 'quality lexicographical data for many languages that otherwise would not or could not exist,' a set of tools that allow to break lexicographical data collection into targeted microtasks were used, as described by Benjamin (2015). The microtasks make it possible to collect translations of a set of words in the target language, to suggest synonyms, to provide inflectional information, examples of usage, and even definitions. In the case of definitions, a term in the target language is provided with the definition of its English translation equivalent, which has been extracted from *Princeton WordNet*. Contributors must write a definition in their own language (which may be a translation of the *WordNet* definition or not). Through a game based on a point-earning system, the next contributors are encouraged to give an improved definition or to vote for an alternative definition proposed by another participant.

These microtasks, which are sometimes gamified on Facebook or smartphones apps, are presented in the public interface, which has been constantly upgraded since the outset of the project. In another paper, Benjamin (2016) looks back at the initial phase of the project, which started two decades earlier: in December 1994, 'the same week as the release of Netscape 1.0,' thirty Swahili speakers who were connected to the Internet were asked to translate English word lists into their language, with the results to be compiled into a static file shared on a Gopher server. There was an intermediary stage between the original phase and the current technological platform: an interface consisting in a form with fields for the words, their part of speech, their definition in Swahili etc. For any word, contributors could edit any field and the dictionary editor, who was notified automatically, could accept or reject the contribution, or modify an entry in turn. This was an early instance of a system implementing the principles of a wiki, which was structured as a database with centralized editorial control. Looking at how the project began reveals that a distributed linguistic work scheme was already in place, using whatever means were at hand. Whether this can be considered as early crowdsourcing or simply as a set of tools facilitating remote communication between linguists/lexicographers and informants is hard to tell. In his 2015 article, Benjamin did describe his system as *crowdsourcing*, but in 2016, he wrote that 'the project has always been conceived as *collaborative* but controlled' (our emphasis). This is yet another example

of an alternate – or hesitant – use of the concepts of collaboration and crowdsourcing.

More recently, Harrison et al. (2019) describe a project for the compiling of talking dictionaries of Zapotec languages that relies on a high level of collaboration between linguists, undergraduate students, technical experts and many Zapotec speakers who actively participate in the design of the dictionary. The collaboration takes place both on site (in person), and remotely, via an online multimedia platform which was developed as part of the project and designed both for browsing the dictionary and for feeding it. For example, the pronunciation of words can be recorded during field surveys or lexicography workshops related to the project. The recordings can also be done remotely and uploaded onto the platform by Internet users. The headword list is established using predefined word lists, legacy sources, and existing teaching material, among other things. Additional words can be collected during thematic conversations or through photo elicitation techniques. The photos, which are extracted from crowdsourced and free naturalist sites, are also used to illustrate entries.

This project is interesting for three reasons. First of all, the speakers volunteering to participate in the dictionary have the same ideological motivating force as the initiators of the project, which they themselves describe as linguistic activism. The authors insist that the methodology – and not only the final product – is central to this project, and that 'the collaborative practices as well as the resulting resources can be interventions in contexts where discrimination and detrimental linguistic ideologies conspire to silence languages.' The issues at stake are, on the one hand, to gain recognition for a language and a culture and, on the other hand, to participate in a revitalization of that language. Everything is done, through collaboration, to strengthen local communities during the recording sessions: e.g. intergenerational sharing of linguistic knowledge, or special interest in diatopic variation from one village to another. There is also online bonding – the platform is linked to social media (e.g. *Twitter* and *Facebook*, where community members communicate in their own language) –, which allows members of the diaspora to reconnect with members of indigenous communities. Secondly, the notion of *prosumer*, which is put forward more often in a theoretical than in an actual way, is embodied in that project in a very concrete manner by the collaborating speakers: the participants actively contribute to a dictionary that they can use, that reflects their culture and that belongs to them. It has been planned from the very beginning of the project that any kind of output will be placed under a free license and any contributing author is systematically credited. Thirdly, just like the Kamusi project described by Benjamin, this project is based on a mix

of traditional approaches to field linguistics and participatory knowledge production via crowdsourcing or collaboration. This hybrid approach shows that volunteers, depending on their motivation, can collaborate with professionals and not only work for them. It also shows that collaboration and crowdsourcing tasks –which can be used jointly– can be specifically tailored to meet the needs of a project and fit into the project's workflow. Finally, it demonstrates that the participatory process can take place outside of wikis and that crowdsourcing can take place outside of the leading platforms.

7 From user feedback to open dictionaries

User feedback did exist before the Internet era in the form of occasional postal mailings sent by users, most of the time to question the presence or absence of a word in the headword list. It is also at the heart of the notion of 'simultaneous feedback' developed by De Schryver and Prinsloo (2000), who believe it should take place throughout the whole dictionary writing process.

Since dictionaries started going online, their users have often been invited to submit comments, in the same way as some online newspapers offer their readers the opportunity to write comments at the bottom of the articles. According to Rundell (2017), this type of feature does not aim to collect users' linguistic knowledge, but to increase user engagement: the more time a user spends on a website, the more income it generates. Some other dictionaries encourage users to contribute in a more precise manner, for example by submitting suggestions of words to be added to the headword list. The *Macmillan Open Dictionary* takes it one step further by asking contributors to submit new words or meanings and to write the corresponding definitions. Once they have been validated by the Macmillan lexicographers (provided they do not contain offensive content and there is evidence showing their use), the contributions get published, without the definitions having to be rewritten in accordance with the dictionary's defining style. Originally designed as a separate lexicon, the crowdsourced open dictionary is now part and parcel of the *Macmillan English Dictionary*. Entries submitted by contributors are clearly indicated as originating from the open dictionary (the contributor's pseudonym, location and date of submission are mentioned) but it is worth mentioning that they are accessible via the same search bar as entries from the 'regular' dictionary. In recently added entries (June 2020), we can find common vocabulary (e.g. *dogsitting* and *misbelief*, which were first recorded in the OED in November 2010 and June 2002), specialized terms (e.g. *symbiont*, 'one of the two organisms involved in symbiosis'), formal neologisms (e.g. *maskne* 'skin irritation

and spots caused by wearing a face mask,' which appeared in *Urban Dictionary* in April 2020 and in *Wiktionary* in July 2020) or semantic neologisms related to current events (e.g. *air bridge* 'a travel arrangement between two countries in which the global outbreak of a disease is under control'). This confirms the ability of the crowds to detect formal and semantic neology, but also to write definitions that are considered, if not perfect, at least acceptable.

8 Ethics

The integration of crowdsourcing and collaborative participation into the lexicographic process is not systematic yet, but some significant milestones have already been set through a variety of projects. Nonetheless, several methodology-related questions remain unanswered. For example: how can the crowds be encouraged to participate in tasks that do not sound very attractive to start with? How should the data collected be assessed? In the case of microtasking, all these problems are closely linked and also raise the question of ethics: since there is paid labour involved and since the rationale behind microtasking is originally to cut the cost, it may be tempting for some to resort to predatory practices (referred to as 'click servitude', 'crowdsploitation' or 'digital slavery') – where should the limit be set? The lack of a national – let alone international – legal framework for online work makes it all the more necessary to reflect upon ethical issues (both from a legal and a moral perspective), even if this goes beyond a purely scientific approach. As pointed out by Bederson and Quinn (2011), it is the responsibility of the designer of the microtasks to establish good practices before any irreversible social damage is done due to wrong technological choices.

8.1 Motivation and remuneration

There are many different reasons why amateurs contribute to a project. Whether these reasons are on the ideological or the utilitarian side, all contributors pursue either a common interest or an individual goal – to name but a few: pursuing a hobby, finding intellectual satisfaction, achieving fame, asserting one's identity, reinforcing a sense of belonging to a community, producing open-source commons, or else acquiring new skills. In the case of annotations carried out in the form of paid microwork, the main motivation remains money (although there might be secondary motivations). Talking about a WSI task, Rumshisky et al. (2012) state – rather bluntly – that, while restricting participation to United States microworkers (i.e. banning Indian contributors) may enhance the quality of annotations, it also requires pay increase, with-

out which Internet users will show little or no interest in the proposed microtasks. Is it legal and desirable to discriminate potential participants on the basis of their origin (or geolocation) without even assessing their competence? For a given task, which amount can be considered fair remuneration, based on duration and the skills required? Can the 'right' compensation be universal or should it be indexed to the cost of living? Under which conditions should remuneration be denied to a microworker?

Entertainment might yet be another motivation, especially by means of Games With A Purpose (GWAP). This consists in designing a system for collecting or annotating data in the form of an online game. *Phrase Detective* (Chamberlain et al., 2009), for example, is a game designed to anaphorically annotate a corpus. However, since hardly any Internet user found anaphora resolution particularly entertaining, a system of rewards in the form of vouchers sent to the highest-scoring players had to be added to the initial version (Poesio et al., 2015). According to Jurgens and Navigli (2014), most GWAPs consist of a text-based interface that makes the game look too similar to a traditional annotation task. As a consequence, the authors suggest developing video games with a graphic design close to the one gamers are familiar with. They get better results with *Puzzle Racer* than with a microwork platform, and at a lower cost (75% less). Those results, however, must be put into perspective for two reasons. First, since participation is ensured by the recruitment of students paid by vouchers, the attractiveness of the game cannot be genuinely assessed. Second, the financial cost of developing the game is nil: it also has to do with student involvement since it was developed as part of a Java course. In comparison, the budget allocated to the salaries of the developers of *Phrase Detective* amounted to £60,000, with vouchers representing an additional budget of £18,000 (Poesio et al., 2015). Getting computer science students to program a GWAP is not so much of an issue. Nor is getting linguistics students to participate in an annotation project highly problematic; it is in fact quite the opposite – it can be very instructive. But how much free work can one reasonably demand from students?

8.2 Quality control: Data evaluation vs workers evaluation

There are several ways to evaluate the data obtained through crowdsourcing, including microworking. This section focuses on the two main methods used for the evaluation of linguistic annotations:[2] (i) the com-

[2] The task-based evaluation of the impact of data on the performance of a system is beyond the scope of this chapter, as it only indirectly – and not intrinsically – evaluates the quality of input data.

parison to a gold standard and (ii) the measuring of an inter-annotator agreement. The development of a gold standard, which is used in particular for the evaluation of machine learning systems requires manual work carried out by experts. As a consequence, it can only be used on a small scale (because of the cost of human experts), which raises the question of the representativeness of the sample thus annotated. Moreover, the dataset produced can hardly be used for any other task or any other type of data than the ones initially targeted (Kilgarriff, 1997). This leaves the inter-annotator agreement, which measures the degree of consensus among raters for a given annotation. There are several measures, including Cohen's kappa, which evaluates the agreement between two annotators, and Fleiss's kappa, which is used for a greater number of annotators.[3] These measures are particularly well suited to the evaluation of annotation by crowdsourcing, which relies on annotation redundancy and consensus building. However, the fact that this measure should only be used as a negative indication is often overlooked: a set of annotations (which is evaluated as a whole) that shows low agreement has to be blamed on unreliable annotations or poorly defined annotation tasks. But the reverse is not necessarily true: high agreement only signals homogeneous annotations, not necessarily quality annotations. The agreement can also be calculated locally, for each annotated unit. Rather than trying to achieve high agreement at all costs (sometimes by distorting the annotation task), Aroyo and Welty (2013) consider that 'annotator disagreement is not noise, but signal; it is not a problem to be overcome, rather it is a source of information.' In the case of WSD, for example, this signal can be used automatically or manually to modify the sense inventory. Chklovski and Mihalcea (2003) rely on web-annotators disagreement to detect sense inventories that might be too fine-grained and to automatically derive coarser-grained inventories from them. For the very same task, Čibej et al. (2015) suggest having a rough draft of sense division drawn up by a lexicographer before requesting the annotators to match up the occurrences with the different senses inventoried. Disagreements may 'alert the lexicographer to an overly coarse sense division or even to an overlooked (sub)sense'.

Evaluating a set of annotations is a complex task that raises methodological questions. The questions raised by the individual assessment of online microworkers could also be considered as methodological considerations as long as what is involved is the discarding of their annotations (i.e. the fact of not using them) when, for some reason, they are deemed unreliable. When it comes to refusing to pay a microworker on such

[3] The theoretical underpinnings and methodological issues raised by these measures are beyond the scope of this chapter. See for example (Artstein and Poesio, 2008) for more detail.

Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography*.
London: Bloomsbury Academic, pp. 343–358

grounds, the question becomes an ethical issue. Problematic workers may fall into two categories: those who are under-qualified for a given task, and deliberate scammers. The first category seems simple to handle and is based on transparent requestor/worker communication. A dataset is used to test the worker at the very beginning, before the actual annotation process, and to discard him/her if he/she does not pass the test. Dealing with malicious workers, i.e. those who try to get paid as quickly as possible by providing the first answer that comes to mind, is more delicate. They may in fact provide correct answers for the initial test and then proceed to cheat. It is possible to introduce occasional questions from a gold standard throughout the annotation process, or to measure the intra-annotator agreement by presenting the same worker with the same item to be annotated several times, at various intervals, in order to test his/her annotation consistency. More often than not, the agreement score is being measured to detect workers who systematically deviate from the others. Dismissing a worker who is too often deviant, i.e. basing one's decision on the 'wisdom of crowds' concept, however tempting, is quite unfair: the majority may be wrong while the individual may be right. Even though the studies led by Snow et al. (2008), which are often cited, show that an NLP system trained on the annotations of several naive annotators obtains better results in several semantic tasks than a system trained on those of a single expert, the findings of Murray and Green (2004), who show that inter-rater agreement is correlated with a homogeneous – and not a high – level of competence among annotators should not be overlooked. Adding the annotations produced by an expert (which are supposed to be quality ones) to those produced by a group of naive people causes the agreement to drop (although, one can imagine this does increase the overall quality of the annotations). So, if, for some reason, it suddenly occurred to a professional lexicographer to participate in a WSD task via a microworking platform, he/she would potentially be detected as a spam worker and would thus be denied payment. Obviously, such a worst-case scenario is not meant to question the need to detect fraudulent behavior, nor the need for procedures proposed to reach that goal. More specifically, it aims to make the case, first and foremost, of the necessary human supervision of decision algorithms.

9 Conclusion

In the first edition of this book published in 2003, Adam Kilgarriff wrote: 'Quite what the role of lexicographer will be, in ten years' time, is far from clear, but I am confident that the role of the corpus will grow, with the line between dictionary and corpus blurring, and the lexicographer operating at the interface.' (see Chapter 7). 2003 was also the

year *Wiktionary* was launched (three years after *Urban Dictionary*) and coincides with the emergence of crowdsourcing. What has happened since then in lexicography regarding the corpus/dictionary continuum and the changing role of the lexicographer has proved him right. Another significant change in the lexicographic process is definitely crowdsourcing and collaborative publishing, which Rundell (2017) sees as an opportunity rather than a threat for professional lexicography. He clearly states that it would be 'foolish to ignore [their] potential': with the right guidance, amateurs can make significant contributions to the design of dictionaries. In his vision of lexicography (which is compatible with Kilgarriff's), the dictionary-making process can be thought of as the division of labour between three participants: lexicographers, machines and volunteer amateurs. Since they each have different assets, the challenge is to find the most efficient configuration for each task to be performed. Some of the configurations that have already been tried out have been listed in this chapter. In the context of a corpus lexicography project, microworking may either be used *together* with NLP tools, or *instead of* them. In the context of what could be named 'field lexicography,' collaborative and/or crowdsourced platforms allow online contributors, considered as informants, to participate in lexical acquisition tasks, or to perform more complex tasks such as the writing of definitions. The latter approach allows the compiling of dictionaries for languages with few or no corpora and tools, which would not have been created otherwise. Finally, *open dictionaries* provide a wide range of additional knowledge overlooked by traditional dictionaries (linguistic knowledge such as regional variations or encyclopaedic knowledge related to specialized fields or subcultures) which allows them to increase their coverage and their receptiveness to lexical innovations.

In addition to the necessary optimization of the distribution of the tasks among computer systems, naive people and lexicographers described by Rundell, the desired efficiency probably also depends on the mutual satisfaction of volunteer workers, publishing houses and dictionary users. Whether crowdsourcing and collaborative knowledge production are about to become the next 'revolution' in lexicography is hard to tell at this point. In the near future, publishers may see it as yet a new way to cut the production costs. May such renewed processes also leave room for further innovation by lexicographers and allow users to gain access to ever-improving dictionaries.

Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography*.
London: Bloomsbury Academic, pp. 343–358

References

Aroyo, L. and Welty, C. (2013). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *ACM Web Science*.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bederson, B. B. and Quinn, A. J. (2011). Web Workers Unite! Addressing Challenges of Online Laborers. In *Proceedings of the International Conference on Human Factors in Computing Systems*, pages 97–106, Vancouver.

Benjamin, M. (2015). Crowdsourcing microdata for cost-effective and reliable lexicography. In *Proceedings of the 9th ASIALEX Conference*, Hong Kong.

Benjamin, M. (2016). Lexicography without lexicographers: Crowdsourcing and the compilation of a multilingual dictionary. Available at `https://www.elexicography.eu/wp-content/uploads/2016/03/Benjamin_lexicography-without-lexicographers.pdf`.

Berg, D., Gönnet, G., and Tompa, F. (1988). The New Oxford English Dictionary Project at the University of Waterloo. Technical Report OED-88-01, Centre for the New Oxford English Dictionary, University of Waterloo.

Brabham, D. C. (2013). *Crowdsourcing*. MIT Press, Cambridge.

Chamberlain, J., Kruschwitz, U., and Poesio, M. (2009). Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In *Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 57–62, Singapore.

Chklovski, T. and Mihalcea, R. (2003). Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. In *Proceedings of RANLP 2003*, Borovets.

Čibej, J., Fišer, D., and Kosem, I. (2015). The role of crowdsourcing in lexicography. In *Proceedings of eLex 2015*, pages 70–83, Herstmonceux.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., and Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word-senses. In *Proceedings of eLex 2013*, pages 49–65, Tallinn.

Damaso, J. (2005). The new Populist Dictionary: A Computer-mediated, Ethnographic Case Study of an Online, Collaboratively-authored English Slang Dictionary. MA dissertation, Queen Mary, University of London.

De Schryver, G.-M. and Prinsloo, D. J. (2000). Dictionary-making process with 'simultaneous feedback' from the target users to the compilers. In *Proceedings EURALEX 2000*, pages 197–209, Stuttgart.

Estellés-Arolas, E., Navarro-Giner, R., and González-Ladrón-de Guevara, F. (2015). Crowdsourcing fundamentals: Definition and typology. In Garrigos-Simon, F. J., Gil-Pechuán, I., and Estelles-Miguel, S., editors, *Advances in Crowdsourcing*, pages 33–48. Springer International Publishing.

Fišer, D. and Ljubešić, N. (2018). Distributional modelling for semantic shift detection. *International Journal of Lexicography*, 32(2):163–183.

Gao, Y. (2012). Online English Dictionaries: Friend or Foe. In *Proceedings EURALEX 2012*, pages 422–433, Oslo.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop*, pages 67–71, Edinburgh.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1489–1501, Berlin.

Hanks, P. (2012). Corpus Evidence and Electronic Lexicography. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*. Oxford University Press, Oxford.

Harrison, K. D., Lillehaugen, B. D., Fahringer, J., and Lopez, F. H. (2019). Zapotec Language Activism and Talking Dictionaries. In *Proceedings of eLex 2019*, pages 31–50, Sintra.

Howe, J. (2006). The rise of crowdsourcing. *Wired*, 14.06.

Jurgens, D. and Navigli, R. (2014). It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. In *Transactions of the ACL*, number 2, pages 449–464.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Kilgarriff, A. (2005). If dictionaries are free, who will buy them? *Kernerman Dictionary News*, 13:17–19.

Kosem, I., Gantar, P., and Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowdsourcing. In *Proceedings of Elex 2013*, pages 32–48, Tallin.

Kosem, I., Krek, S., Gantar, P., Holdt, Š. A., Čibej, J., and Laskowski, C. (2018). Collocations Dictionary of Modern Slovene. In *Proceedings of the 18th EURALEX Congress*, pages 989–997, Ljubljana.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th EACL Conference*, pages 591–601, Avignon.

Manandhar, S., Klapaftis, I. P., Dligach, D., and Pradhan, S. S. (2010). SemEval-2010 Task 14: Word Sense Induction & Disambiguation. In *Proceedings of the 5th Workshop on Semantic Evaluation*, pages 63–68, Los Angeles.

Meyer, C. M. and Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*, pages 259–291. Oxford University Press, Oxford.

Murray, G. C. and Green, R. (2004). Lexical Knowledge and Human Disagreement on a WSD Task. *Computer Speech & Language*, 18(3):209–222.

Nagao, M., Tsujii, J., Ueda, Y., and Takiyama, M. (1980). An attempt to computerized dictionary data bases. In *Proceedings of COLING 1980*, pages 534–542, Tokyo.

Nesi, H. (2008). Dictionaries in electronic form. In Cowie, A. P., editor, *The Oxford History of English Lexicography*, pages 458–478. Oxford University Press, Oxford.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Luca, D. (2015). Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. In *Proceedings of IJCAI 2015*, pages 4202–4206, Buenos Aires.

Rumshisky, A. (2011). Crowdsourcing Word Sense Definition. In *Proceedings of the Fifth Law Workshop*, pages 74–81, Portland.

Rumshisky, A., Botchan, N., Kushkuley, S., and Pustejovsky, J. (2012). Word Sense Inventories by Non-experts. In *Proceedings of the 8th LREC Conference*, pages 4055–4059, Istanbul.

Howard Jackson (ed), *The Bloomsbury Handbook of Lexicography*.
London: Bloomsbury Academic, pp. 343–358

Rundell, M. (2014). Macmillan English Dictionary: The End of Print? *Slovenščina 2.0*, 2(2):1–14.

Rundell, M. (2017). Dictionaries and crowdsourcing, wikis, and user-generated content. In Hanks, P. and de Schryver, G.-M., editors, *International Handbook of Modern Lexis and Lexicography*. Springer, Berlin, Heidelberg.

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., editors, *A Taste for Corpora. In honour of Sylviane Granger*, pages 257–282. John Benjamins.

Rundell, M. and Stock, P. (1992). The corpus revolution. *English Today*, 30:9–14.

Sajous, F. and Hathout, N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of eLex 2015*, pages 405–426, Herstmonceux.

Sajous, F., Hathout, N., and Josselin-Leray, A. (2019). *Du vin* et *devin* dans le Wiktionnaire : neutralité de point de vue ou neutralité *et* point de vue ? *Études de linguistique appliquée*, 194(2):147–164.

Sajous, F., Josselin-Leray, A., and Hathout, N. (2018). The Complementarity of Crowdsourced Dictionaries and Professional Dictionaries viewed through the Filter of Neology. *Lexis*, 12.

Sajous, F., Josselin-Leray, A., and Hathout, N. (2020). Les domaines de spécialité dans les dictionnaires généraux : le lexique de l'informatique analysé par les foules et par les professionnels... de la lexicographie. *Neologica*, 14:83–107.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and Fast—but is it good?: Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 8th EMNLP Conference*, pages 254–263, Morristown.

Trap-Jensen, L. (2018). Lexicography between NLP and Linguistics: Aspects of Theory and Practice. In *Proceedings of EURALEX 2018*, pages 25–37, Ljubljana.