Crowdsourcing workflows in lexicography

Darja Fišer and Jaka Čibej

Abstract

The success of a crowdsourcing campaign depends on a number of factors, e.g. an effective workflow, the funding available, the technological framework for crowdsourcing, the type of crowdsourcer motivation, and the type and volume of the data to be processed. Before embarking on a project it is therefore imperative to analyse its needs and plan the implementation of crowdsourcing that best fits the specific circumstances, to ensure the feasibility of the campaign and good results. In this paper we propose a general crowdsourcing workflow for lexicographic projects that can then be tailored to various scenarios. We also provide an overview of the most popular crowdsourcing platforms and discuss the criteria to be taken into account when selecting the one used for a specific lexicographic project.

Keywords: crowdsourcing, workflow, dictionary construction, crowdsourcing platforms

1 INTRODUCTION

Crowdsourcing has a lot of potential in contemporary lexicographic projects, especially as a means to post-process automatically extracted data and facilitate the work of lexicographers. Although crowdsourcing has not yet been thoroughly tested on large-scale lexicographic projects, a number of related projects have proved that it can be both sufficiently accurate and effective (cf. Klubička and Ljubešić 2014; Fišer et al. 2015; Kosem et al. 2013b). These encouraging results indicate that the power of the crowd could also be harnessed in the field of lexicography. However, each crowdsourcing campaign needs to take into account a number of external factors such as the budget and time available, the amount and type of data that needs to be processed, as well as the pool and type of crowdsourcers that can be recruited. In this paper we propose a general workflow for lexicographic projects, each step of which can be tailored to specific project circumstances. We then describe a set of crowdsourcing scenarios for the most common lexicographic project types, highlight the key principles that need to be taken into account and present the customized workflow for each of these. Finally, we give an overview of the most popular crowdsourcing platforms and present the criteria for choosing the best one for the lexicographic project at hand.

2 CROWDSOURCING WORKFLOW IN LEXICOGRAPHIC PROJECTS

In this section, we propose a general crowdsourcing workflow that can be used in various phases of corpus-based lexicographic projects. Our approach is modular and can therefore be adapted according to the needs of the project at hand. The order of the stages can be changed, some can be done in parallel or even left out, but it is important to at least consider the stages we recommend and address the issues each of them raise, as crowdsourcing is a complex, time-consuming and potentially costly procedure that cannot yield useful results without careful planning and task design.

Before deciding on a crowdsourcing campaign, an estimate of the required investment should be made with respect to the time, money and personnel required, as the campaign should not take up more time and financial and/or human resources than conventional annotation methods. An important advantage of including crowdsourcing from the very beginning of dictionary project planning is the fact that the initial input in the preparation of an appropriate crowdsourcing environment pays off in the long run: crowdsourcing can be used in numerous phases of dictionary construction, microtasks can be

designed according to the same principles, and data can be annotated and processed using the same platform.

We describe the individual stages of the crowdsourcing workflow in the following sections.

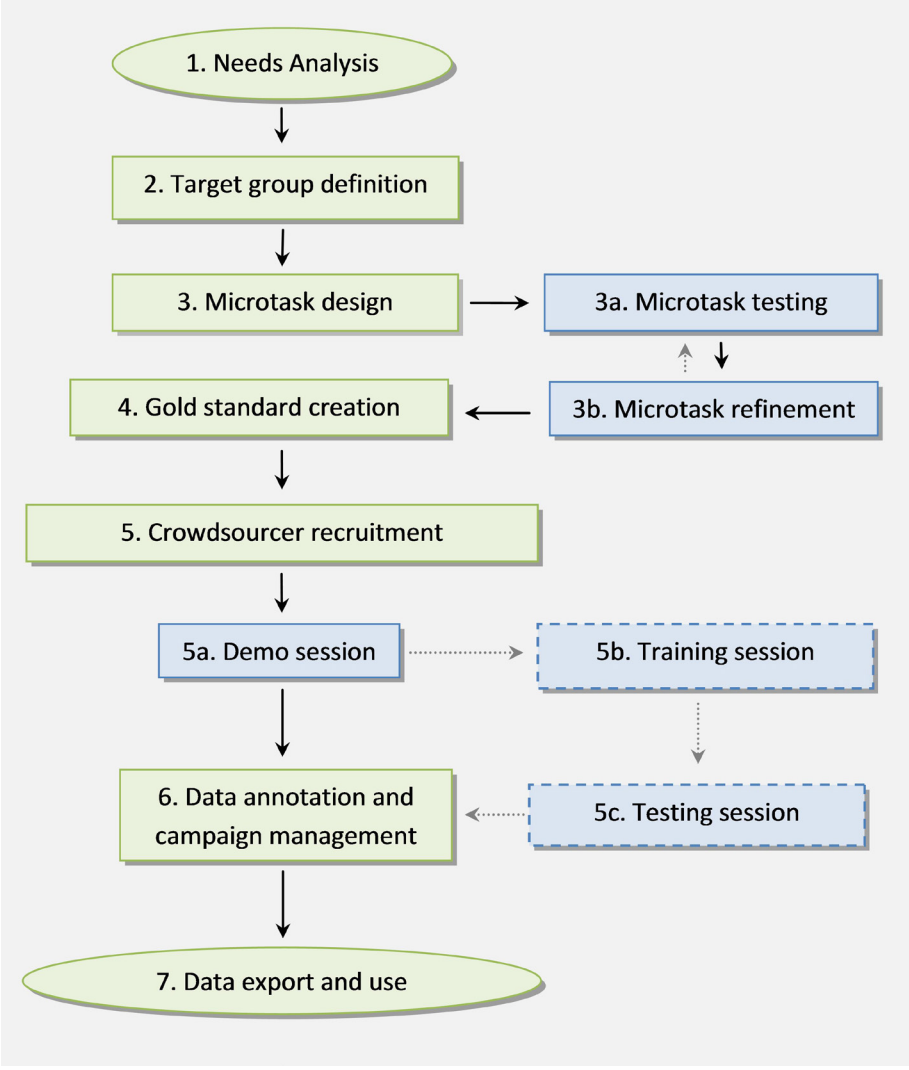

Figure 1: Crowdsourcing workflow for lexicography. Green-coloured boxes represent the main stages, while blue-coloured ones are the subphases. Dashed boxes and arrows represent optional stages which can be omitted in small-scale, low-budget campaigns.

2.1 Needs analysis

The first step of each crowdsourcing campaign requires a thorough needs analysis. Several things need to be determined: the goals and expectations of the campaign, the quantity of the data to be processed, the purpose for which it is to be used, as well as its format and availability. With dictionary projects, in which crowdsourcing can be used in different phases, it is recommended to analyse the needs of each phase and design the workflow, platform and timeline of the crowdsourcing campaign in such a way that it ensures compatibility of input data and software throughout the entire project.

2.2 Crowdsourcer profile

Once the needs have been analysed it is necessary to determine the required crowdsourcer profile, as tasks can vary in complexity and require different skills. The problem at hand may be suitable for the general public without any specialized linguistic or lexicographic knowledge, or it may require a certain degree of expertise and can only be solved effectively by language students or expert lexicographers.

3.3 Microtask design, testing and refinement

The most important and difficult part of crowdsourcing is microtask design. Microtasks should be one-dimensional questions with concise instructions and suited to the target crowdsourcer profile.

For instance, tasks aimed at the general public should not contain terminology or complex structures, which should be replaced with practical examples (e.g. the question "*Which meaning best corresponds to the use of the word in the phrase contained in the example?*" can be simplified to "*What is the meaning of the underlined word in the sentence below?*"). It is very important not to design microtasks in such a way that they yield unreliable results. This is especially problematic with multi-dimensional questions, as crowdsourcers will not be able to answer them accurately (e.g. the question "*Is the collocation below suitable to be included in a dictionary?*" can be divided into two parts: 1. "*Is the collocation below correctly extracted from the corpus?*" and 2. "*Does the collocation below fit into a learning dictionary?*").

The designed microtasks need to be tested in a pilot study in order to check their effectiveness, determine potential incongruences or mistakes, and eliminate all identified shortcomings. If a microtask turns out to be too complex for the chosen crowdsourcer profile, it needs to be adapted or reassigned to crowdsourcers with more expertise in the field.

2.4 Gold standard creation

A certain number of microtasks needs to be annotated by experts to create a gold standard that is later used to ensure the quality of the crowdsourced results. The dataset should be as representative of the entire set of microtasks as possible, both in terms of size and complexity.

2.5 Crowdsourcer recruitment and training

After microtasks have been designed and the gold standard created, it is time for crowdsourcer recruitment and training. The crowdsourcing initiator usually holds a **demo session**, either live or, most often, as a presentation or demo video that is made available on the project website. The demo session introduces crowdsourcers to the annotation process. This is followed by a **training session**, during which crowdsourcers solve tasks under the supervision of an expert who provides advice or further explanation. Alternatively, the training session can be held online with automatic feedback for every solved task. The final recruitment step is the **testing session**, during which crowdsourcers solve tasks independently and are recruited if they pass a given accuracy threshold. With low-budget projects, training and testing sessions are often combined with the main part of the campaign, while the unreliable results/crowdsourcers are excluded.

2.6 Data annotation and campaign management

This is the main stage of every crowdsourcing campaign, during which the recruited crowdsourcers solve the microtasks provided by the initiator. The initiator needs to monitor the campaign and decide whether any additional fine-tuning is necessary, e.g. whether the set of microtasks needs to be expanded, the crowdsourcers are motivated enough to provide a consistent flow of answers, and so on.

2.7 Data export and use

The final stage involves exporting the crowdsourced data into an appropriate format for further use in the project (e.g. for algorithm training or inclusion in a dictionary). The crowdsourcing platform should allow the data to be exported at any point of the crowdsourcing campaign, as checking whether interim results are meeting the expectations of the project is crucial for good campaign management.

3 TYPES OF LEXICOGRAPHIC PROJECTS

In this section, we present potential scenarios of implementing crowdsourcing into various types of lexicographic projects. As already emphasised, the flow of the crowdsourcing campaign depends a great deal on funding. Funding is directly related to the range and timeframe of the crowdsourcing campaign, the project phases in which crowdsourcing will be used, the number of microtask types designed, the complexity of the crowdsourcing workflow, the number of recruited crowdsourcers, and the type of motivation used. The more financial resources there are available, the more specialised the applications that can be developed, tested, optimised and finally presented to a wide circle of crowdsourcers. Low-budget projects, on the other hand, require more input when it comes to recruiting and motivating crowdsourcers. However, the social motivation of crowdsourcers can (and should) be used in all scenarios.

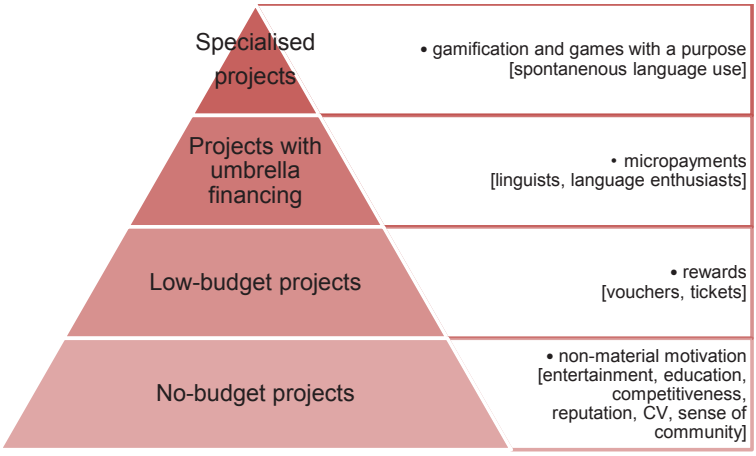

Figure 2: An overview of crowdsourcing scenarios for various types of lexicographic projects.

3.1 Specialised projects

Most specialised projects with full financing can afford tailor-made crowdsourcing applications, most notably games with a purpose (GWAPs). Their entertaining and competitive elements ensure three long-term interest of a wide group of players and spontaneous language use. GWAPs have proved highly successful in a number of related projects (cf. Jurgens and Navigli 2014; Joubert and Lafourcade 2012; Chamberlain et al. 2008). Jurgens and Navigli (2014) found that *Puzzle Racer*, a game that involves players annotating corpus data, achieves the same level of quality as conventional data annotation by experts, while lowering the overall costs by 73% compared to a classical crowdsourcing campaign involving microtasks. A specialised GWAP can be used to collect large quantities of data, can be adapted for different devices and platforms, and allows for the inclusion of different tasks for different profiles and phases of the lexicographic project.

3.2 Projects with umbrella financing

Many contemporary lexicographic projects have no direct funding and are instead realised as one of the non-primary activities of a wider research project or programme. In this scenario, it is recommended to use existing resources and technologies to plan a crowdsourcing campaign in such a way that the results can be directly applicable, not only in the context of the lexicographic project, but also to the main project and any future projects that arise. The resources to develop customized applications are most likely not available, but any of the popular crowdsourcing platforms can be used. The campaign should attract lexicographers, language editors, translators and language enthusiasts who can be paid through micropayments. The number of microtasks, the quantity of the crowdsourced data, and the number of crowdsourcers should correspond to the available financial resources. If necessary, optional phases (e.g. cyclic microtask editing, crowdsourcer training and testing) can be left out of the workflow (see Figure 1).

3.3 Low-budget projects

In low-budget projects it is recommended to invest the majority of the financial resources available in automating data preparation as much as possible, while also significantly simplifying the crowdsourcing workflow. In this scenario, the default

quality control parameters can be used and the majority vote without an expert referee can be used to make final annotation decisions. The crowdsourcers best suited for this scenario are students of linguistics or language enthusiasts, whose work can be rewarded with vouchers, tickets or other material rewards. This approach has already proved to be feasible (El-Haj et al. 2014; Fišer et al. 2015). However, it demands realistic expectations when it comes to crowdsourcer input in terms of time and effort. The crowdsourcers should not be presented with overly ambitious tasks, nor should they be expected to do a significant amount of work in a short period of time – a fact that needs to be taken into consideration when planning the project. Instead, a longer campaign should be foreseen compared to scenarios with funding.

3.4 No-budget projects

In cases when no financial resources are available, crowdsourcing can be implemented in a manner similar to that employed by numerous collaborative lexicographic projects that recruit and motivate crowdsourcers with non-material rewards based on social motivation. Aside from enthusiasts who enjoy contributing to the construction of new language resources, the wider public can also be motivated to join the campaign if offered entertaining tasks or organised competitions (Fišer et al. 2015). In addition, students and graduates can be recruited by offering awards or certificates for participating in the project, which they can use for extra-credit or as a reference to add to their CV.

As with the low-budget scenario, it is crucial to plan a no-budget crowdsourcing campaign as a long-term project. Crowdsourcers should only be given simple tasks, and the project should be relevant for their community. It is also necessary to take into consideration the fact that the crowdsourcers involved are participating out of enthusiasm for the project, which is why it is even more important to keep in touch with them regularly and build a well-connected community.

4 CROWDSOURCING PLATFORM SELECTION

A crowdsourcing platform is an application that allows the crowdsourcing initiator to upload a project containing microtasks that are then solved by the recruited crowdsourcers. In this section, we describe the criteria to follow when selecting an appropriate platform, as well as the process of choosing the platform that is to be used for crowdsourcing the *Dictionary of Modern Slovene Language*.

4.1 Selection criteria

The selection of a suitable platform is one of the first steps to undertake in a crowdsourcing campaign. Several criteria need to be taken into account.

Data format – The platform needs to support uploading different types of microtasks and exporting crowdsourcing results in formats suitable for the needs of the project.

Interface – It is important for the platform to offer a simple, user-friendly interface both for the campaign administrator and the crowdsourcers. The administrator should be able to use the platform to form different types of tasks of varying complexity, to monitor the statistics of data collection and crowdsourcer reliability, to expand the gold standard if necessary (without interrupting the crowdsourcing process), or update the set of microtasks and export preliminary results. The crowdsourcers, on the other hand, should be provided with a simple registration process (e.g. using a Gmail, Twitter or Facebook account), personal data protection, and a comfortable working environment that increases their motivation.

Quality control – It is important to make sure that the platform contains as many quality control measures as possible, e.g. a gold standard, inter-annotator agreement, crowdsourcer consistency, and majority vote. In addition, the platform should allow the administrator to fine-tune the settings that control the inclusion of gold standard microtasks into crowdsourcer tasks, repeating the same microtasks with multiple crowdsourcers, the time restriction for individual tasks, and so on.

Financial aspect – The platform needs to support micropayments if this type of economic motivation is to be used to motivate crowdsourcers. With commercial crowdsourcing platforms that offer campaign hosting, the amount the crowdsourcing initiator needs to transfer depends on the size and complexity of the campaign. The bulk of the resources are used for micropayments (their size is usually determined by the crowdsourcing initiator), while a certain percentage is taken by the platform manager.

Motivational mechanisms – It is advantageous if the platform already incorporates mechanisms for additional crowdsourcer motivation, e.g. a scoring system, hall of fame, automatic notifications when someone beats the current high score, and automatic reminders for crowdsourcers who have been inactive for longer periods of time.

4.2 Overview of crowdsourcing platforms

When selecting a platform for the construction of a dictionary of modern Slovene, we reviewed approximately 150 crowdsourcing platforms between October and November 2014. In the following sections, we list and describe those that are suitable for crowdsourcing linguistic data.

4.2.1 Commercial platforms

The most popular crowdsourcing platform is Amazon Mechanical Turk.[1] Its interface already incorporates mechanisms for quality control, campaign management and micropayment support. The platform also has a large existing pool of registered crowdsourcers. However, they are mostly speakers of larger languages.

Similar examples are Crowdflower[2] and Clickworker.[3] Both offer a number of applications for various fields of data processing (e.g. data categorisation and sentiment analysis). Microtasks can be uploaded in CML, CSS or Javascript, and crowdsourcers can be filtered according to their age, knowledge prerequisites and geolocation.

4.2.2 Open-source platforms

The most notable open-source platform is Crowdcrafting,[4] which recruits volunteer crowdsourcers to contribute to various research projects by solving tasks. The platform is based on PyBossa,[5] an open-access software for creating crowdsourcing projects that can be installed on a local server and is available under the Creative Commons BY-SA 4.0 licence.

Another open-source crowdsourcing tool is sloWCrowd[6] (Tavčar et al. 2012), which is PHP/MySQL-based and was developed for cleaning automatically generated sloWnet synsets but later extended to enable other types of crowdsourcing tasks (Fišer et al. 2015).

1 https://www.mturk.com
2 http://www.crowdflower.com
3 http://www.clickworker.com/
4 http://crowdcrafting.org/
5 http://pybossa.com/
6 http://nl.ijs.si/slowcrowd/

4.3 Platform selection for the construction of the *Dictionary of Modern Slovene*

After reviewing the existing crowdsourcing platforms, we decided to use PyBossa for the crowdsourcing of the *Dictionary of Modern Slovene*. The reasons for this are as follows:

Flexibility – Unlike commercial platforms, PyBossa can be installed on a local server, and its interface can be adapted to the needs and conditions of the project.

Support – As an open-source platform, PyBossa is well supported and constantly developed. It has already been successfully used in numerous projects, and many additional libraries are available to enable more mechanisms for monitoring the results of the crowdsourcing process, and other outcomes.

Financial independence – In case of insufficient funds, paying crowdsourcers with micropayments will not be possible. Commercial platforms do not offer other types of payment (rewards, tickets, etc.). In addition, using an open-source platform will save the commission that needs to be paid to professional crowdsourcing platforms for handling the micropayments.

Logistical reasons – There are a number of technical barriers when dealing with commercial platforms. For instance, Amazon Mechanical Turk requires the crowdsourcing initiator to have a bank account in the US. In addition, the platform would require registration and personal data from every Slovene crowdsourcer, which is very inconvenient. Difficulties would probably arise with micropayments as well, as the spending of public funds is strictly regulated in Slovenia.

Best practice – PyBossa has already been successfully used for crowdsourcing in numerous research projects. The platform's website[7] lists a number of users, e.g. the British Museum, the Swiss Research Institute CERN, and UNITAR.

5 LIMITATIONS OF CROWDSOURCING

Despite the great deal of attention crowdsourcing has recently received among lexicographers, misconceptions and prejudices about it are still common. We address these issues in this section.

To ensure the appropriate role of crowdsourcing in lexicographic projects, it is imperative to recognise its limitations as well as its potential. Crowdsourcing is

7 http://crowdcrafting.org/about

not effective for every type of data, every phase of lexicographic work or every lexicographic project. For instance, it cannot be implemented unless regular campaign management can be guaranteed (designing microtasks, controlling the collected answers, motivating and paying crowdsourcers). Crowdsourcing is also not suitable for open-ended questions or tasks that require subjective answers. It can only be useful when it saves time and/or financial resources for the lexicographic project, despite all the preparation and management that it warrants, while still providing reliable results.

5.1 Amateur lexicographers

Because certain authors are somewhat inconsistent when defining crowdsourcing (cf. Estellés-Arolas and González-Ladrón-de-Guevara 2012), it is often unjustifiably mistaken for – or even equated with – collaborative lexicography. Unlike numerous collaborative projects in which all the work is done by non-experts, a project initiator or manager is always heavily involved in crowdsourcing by preparing data, designing microtasks, controlling quality, ensuring crowdsourcer motivation, and so on. Although some collaborative projects have shown that users can also contribute to useful and widely used dictionary products (Meyer and Gurevych 2012), crowdsourcing as proposed in this paper primarily involves post-processing automatically extracted corpus and lexicon data before actual dictionary construction. Furthermore, user contributions are not immediately published as the content of the dictionary, and the organisation of lexicographic information is still controlled by the lexicographers.

Meyer and Gurevych (2012) pointed out that collaborative projects represent the sum of the opinions of numerous authors, who put considerable effort into improving dictionary entries until a consensus is reached on their structure and content. For this reason, collaborative lexicography in many respects gives results comparable to official lexicographic products. However, its biggest shortcoming is the lack of an effective mechanism to separate mature and high-quality dictionary entries from those that still require improvement. A similar observation was made by Lew (2013), who noted that in certain cases the order of definitions in Wiktionary can be somewhat random, with completely marginal meanings displayed at the top. A similar issue is found in Urban Dictionary, where users can vote to influence the order of definitions, and the most popular definition is not necessarily the most appropriate, but rather one that best reflects the users' ideology or the one they find most entertaining.

In contrast, we envisage crowdsourcing as one of the phases of dictionary construction. First, data is automatically extracted from corpora and other datasets.

The data is then post-processed by crowdsourcers through microtasks and finally used by lexicographers in manual lexicographic work. Crowdsourcing is thus an intermediate link between automatic data processing and manual expert data processing, as it significantly reduces the lexicographer's workload through automatic data extraction and crowdsourcing, while also including a manual approach in processes that still cannot be automated effectively. Although crowdsourcing has not yet been thoroughly tested on large-scale lexicographic work, the results of related projects have proved effective (Klubička and Ljubešić 2014; Fišer et al. 2015; Kosem et al. 2013b) and indicate that it can be successfully implemented in the field of lexicography.

5.2 Reliability of crowdsourced dictionaries

A common misconception about crowdsourced results is that they are unreliable, especially because the pool of crowdsourcers can include non-experts. We emphasise that microtasks should always be designed for a specific crowdsourcer profile and take into account their level of expertise. A well-designed crowdsourcing project will assign more complex tasks to crowdsourcers with more expertise in the field (e.g. students or graduates of linguistics), while only simple tasks will be left to non-experts.

Fišer and Čibej (2015) presented a number of quality control mechanisms, e.g. a gold standard, inter-annotator agreement, majority vote, consistency, and refereeing. These can be used to effectively eliminate those crowdsourcers that provide incorrect or unreliable answers. These quality control measures have already been tested by numerous authors (cf. Rumshisky 2011; Fišer et al. 2015; Klubička and Ljubešić 2014; Fossati et al. 2013), and found to ensure high accuracy of crowdsourcing results that achieve the same level of quality as if the work were done only by experts (Snow et al. 2008).

5.3 Impact of crowdsourcing on lexicography

As a new form of work not yet explicitly covered by legislation, crowdsourcing undoubtedly raises many ethical issues regarding payment, work conditions and authorship recognition. Crowdsourcing platform providers act as employment agencies, but it is the crowdsourcing initiators who determine payment conditions and the work load. Although crowdsourcers are not obligated to accept badly paid tasks, they are often forced to if they want to earn a living. Low

payments and unfair work practices in language resource crowdsourcing have been criticized by several authors (Sabou et al. 2014; Silberman et al. 2010; Lease and Alonso 2014; Felstiner 2011). For example, e.g. Snow et al. (2008) offered a total of 2$ for 7,000 non-expert annotations, and 1$ for 1,500 expert annotations via Amazon Mechanical Turk. It is thus the duty of the coordinators of every lexicographic project to treat crowdsourcers fairly and credit their contributions to the final product. Their pay needs to be taken into account at the very inception of the project, when the budget is determined.

In addition to issues stemming from payment and work conditions, crowdsourcing has also faced accusations that it degrades the profession of lexicographers and linguistics, redirecting their workload to an unqualified (and poorly paid) crowd. We wish to emphasise that the basic concept of crowdsourcing in this context is the rational use of resources: expert lexicographers are spared trivial and routine tasks, and crowdsourcers can contribute to language resource construction as best they can, while at the same time receiving different forms of motivation (monetary or material rewards, gaining experience and references, entertainment, etc.).

Because the misconceptions surrounding crowdsourcing campaigns are not only present among experts, but also in the general public, it is important for the crowdsourcing initiator to form an intelligent strategy for public relations. Communication with the potential crowdsourcers needs to be carried out with great care and respect, and the input expected from the crowd should reflect the type of motivation. For instance, if the crowd receives no monetary remuneration for their work then it should not be presented with overly ambitious tasks. It is also important for the crowdsourcing initiator to keep in touch with the crowdsourcers throughout the campaign, e.g. by informing them about the project workflow, inviting them to project presentations or similar events, and publicly thanking them for their contributions.

6 CONCLUSION

A well-planned crowdsourcing project that observes the key principles of microtask design and management can be of great help in lexicography, as it can handle the post-processing of automatically extracted noisy data in an economical and timely manner, with reliable results. This paper gave a comprehensive and detailed account of the organisational, technical, linguistic as well as financial aspects of successful crowdsourcing for dictionary creation, and proposed a general crowdsourcing workflow as well as several specialised scenarios that take into account various lexicographic project types and circumstances.

Crowdsourcing has already been embraced by language technologies and language resource creation. Recent successful small-scale specialized lexicographic projects have built a firm foundation for crowdsourcing to be included in more complex, large-scale lexicographic projects. The dictionary of modern Slovene is one of the first lexicographic projects that plans on implementing crowdsourcing in its entire workflow, and, as a pioneer project, it will pave the way for future dictionaries and language resources, both in Slovenia and abroad.