Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 111–120
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Text Corpora and the Challenge of Newly Written Languages

Alice Millour*, Karën Fort*†
*Sorbonne Université / STIH, †Université de Lorraine, CNRS, Inria, LORIA
28, rue Serpente 75006 Paris, France, 54000 Nancy, France
alice.millour@etu.sorbonne-universite.fr, karen.fort@sorbonne-universite.fr

Abstract

Text corpora represent the foundation on which most natural language processing systems rely. However, for many languages, collecting or building a text corpus of a sufficient size still remains a complex issue, especially for corpora that are accessible and distributed under a clear license allowing modification (such as annotation) and further resharing. In this paper, we review the sources of text corpora usually called upon to fill the gap in low-resource contexts, and how crowdsourcing has been used to build linguistic resources. Then, we present our own experiments with crowdsourcing text corpora and an analysis of the obstacles we encountered. Although the results obtained in terms of participation are still unsatisfactory, we advocate that the effort towards a greater involvement of the speakers should be pursued, especially when the language of interest is newly written.
Keywords: text corpora, dialectal variants, spelling, crowdsourcing

1. Introduction

Speakers from various linguistic communities are increasingly writing their languages (Outinoff, 2012), and the Internet is increasingly multilingual.[1]

Many of these languages are less-resourced: these linguistic productions are not sufficiently documented, while there is an urge to provide tools that sustain their digital use. What is more, when a language is not standardized, we need to build adapted resources and tools that embrace its diversity.

Including these linguistic productions into natural language processing (NLP) pipelines hence requires efforts on two complementary fronts: (i) collecting or building resources that represent the use of the language, (ii) developing tools which can cope with the variation mechanisms.

In all cases, the very first resource that is needed for further processing is a text corpus.

After presenting the challenges related to the processing of oral languages when they come to be written, we introduce in Section 3. the existing multilingual sources that are commonly used to collect corpora, as well as their main shortcomings.

In Section 4., we show how crowdsourcing has been used in the past to involve the members of linguistic communities into collaboratively building resources for their languages. We argue that this method is all the more reasonable when it comes to collecting meaningful data for non-standardized languages.

In Section 5., we present the existing initiatives to crowdsource text corpora. Based on our own experiments and on the result of a survey regarding the digital use of a non-standardized language, we explain why collecting this particular type of resource is challenging.

2. The Need for Text Corpora

The rise in use of SMS, online chat and of social media in general has created a new space of expression for an increasing number of speakers (see for instance, the studies on specific languages carried by Rivron (2012) or Soria et al. (2018)). Although linguistic communities are being threatened all over the world, this represents a valuable opportunity to observe, document and equip with appropriate tools an increasing number of languages. These new spaces of written conversation have been taken over by linguistic communities which practice had been mainly oral until then (van Esch et al., 2019). When no orthography has been defined for a given language, or when one (or various) conventions exist but are not consistently used by the speakers, spellings may vary from one speaker to another. Indeed, when the spellings are not standardized by an arbitrary convention, speakers may transcribe the language based on how they *speak* it.

In fact, standardizing spelling does not confine to defining the orthography, as it usually also acts as a unification process of potential linguistic variants towards a sole written form. By contrast, the absence of such a standardization process authorizes the raw transcription of a multitude of linguistic variants of a given language. These variants being transcribed according to the spelling habits and linguistic backgrounds of each speaker, the Internet, especially in its conversational nature, has become a breeding ground for linguistic diversity expression and observation.

Situations of spelling variations observed on the Internet are documented in diverse linguistic contexts such as the ones of:

- The Zapotec and Chatino communities in Oaxaca, Mexico, as detailed in (Lillehaugen, 2016) in the context of the *Voces del Valle* program. During this program, speakers were encouraged to write tweets in their languages. They were provided with spelling guidance they were not compelled to follow. As stated by the author: "*The result was that, for the most part, the writers were non-systematic in their spelling decisions—but they were writing*".

- Some of the communities speaking Tibetan dialects outside China, which develop a "*written form based*

[1] See, for instance, the reports provided by `w3tech` such as `https://w3techs.com/technologies/history_overview/content_language/ms/y`.

on the spoken language" independently of the Classical Literary Tibetan (Tournadre, 2014).

- The Eton ethnic group, in Cameroon, about which Rivron (2012) observes that Internet is the support of "*the extension of a mother tongue outside its habitual context and uses, and the correlated development of its graphic system*".
- Speakers of Javanese dialects who "*have their own way of writing down the words they use according to the pronunciation they understand*", regardless of the official spelling. Each dialect developing its own spelling, a dialect that was "*originally only recognizable through its oral narratives (pronunciation) is now easily recognizable through the spelling used in social media*" (Fauzi and Puspitorini, 2018).
- Communities using Arabizi to transcribe Arabic online: as reported by Tobaili et al. (2019), Arabizi allows multiple mappings between Arabic and Roman alphanumerical characters, and thus makes apparent dialectal variations usually hidden in the traditional writing.
- Communities transliterating Indian dialects with Roman alphabet without observing systematic conventions for transliteration (Shekhar et al., 2018).
- Regional European languages such as Alsatian, a continuum of Alemannic dialects, for which a great diversity of spellings is reported (Millour and Fort, 2019) even though a flexible spelling system, Orthal (Crévenat-Werner and Zeidler, 2008), has been developed.

In the following, we will refer to these proteiform languages as "multi-variant". The variation observed is indeed the result of (at least) two simultaneous mechanisms: the dialectal and scriptural variations. Both of these degrees of freedom may be impacted by the usual dimensions for variation (diachronic, diatopic, diastratic, and diamesic).
From a NLP perspective, these linguistic productions represent a challenge. In fact, they push us to deal with the issues that variation processes imply, either because these productions account for most of the written existence of a non-standardized language, or because they diverge from a standard language in an undeterministic fashion. Yet, for the endangered languages there is an urge to develop tools that match the actual linguistic practice of its end-users to sustain their digital use.
Even though less-resourced languages benefit from the current trends in NLP which tend towards less supervision (see, for instance (Lample et al., 2017; Grave et al., 2018)) and seek higher robustness to variation, processing technologies still highly rely on the availability of text corpora.

3. Existing Sources Used for Corpus Collection

Although there exist sources of text corpus readily available for numerous languages, these "opportunistic" corpora (McEnery and Hardie, 2011) present several shortcomings, including:

- an insufficient coverage to constitute the basis for further linguistic resources developments. These corpora are unlikely to be balanced in terms of representativeness of the existing practices.
- their nature and license sometimes require operations that result in a loss of information such as the metadata necessary to identify the languages or the structure of the document.
- using them requires additional linguistic resources (to perform language identification, for instance).

In the following, we first present the Wikipedia project, which, with 306 active Wikipedias distributed under Creative Commons licenses, undoubtedly provides the largest freely available multilingual corpus.
Second, we present how the Web can more generally be used as a source of text corpora. We focus on describing how Web crawling has been used to gather corpora for less-resourced languages, and briefly comment on the use of social networks-based corpora.

3.1. Wikipedia as a Corpus

Wikipedia is an online collaborative multilingual encyclopedia supported by the WIKIMEDIA FOUNDATION, a non-profit organization.
Along with providing structured information from which lexical semantic resources or ontologies can be derived, Wikipedia is an easily accessible source of text corpora widely used in the NLP community, and from which both well- and less-resourced languages benefit.

Its popularity and its collaborative structure make it the most natural environment to foster collaborative text production. We discuss in this section to which extent Wikipedia represents a valuable source of text corpora for less-resourced and non-standardized languages, in terms of further NLP processing.
After a short introduction on the size and quality of the existing Wikipedias, we describe how the issue of language identification and the purpose of Wikipedia prevent it to be the most appropriate virtual place to host dialectal and scriptural diversity.

3.1.1. Size and Quality of Wikipedias

There exist 306 active Wikipedias[2], 16 of them showcasing more than 1 million articles, 62 more than 100,000, 147 more than 10,000, and 81 between 1,000 and 10,000.
Even though observing the size of the Wikipedia in terms of article count gives a useful overview of the linguistic diversity of the project, size is not the best indicator to get a sense of the amount of quality data available in each Wikipedia. Instead, the `Depth` indicator[3] has been defined by WIKIMEDIA to get an estimate of the quality of a given Wikipedia based on the number of articles, but also edits, and proportion of "non-article" pages such as user

[2] See `https://meta.wikimedia.org/wiki/List_of_Wikipedias`, as of January 2020.
[3] See `https://meta.wikimedia.org/wiki/Wikipedia_article_depth`

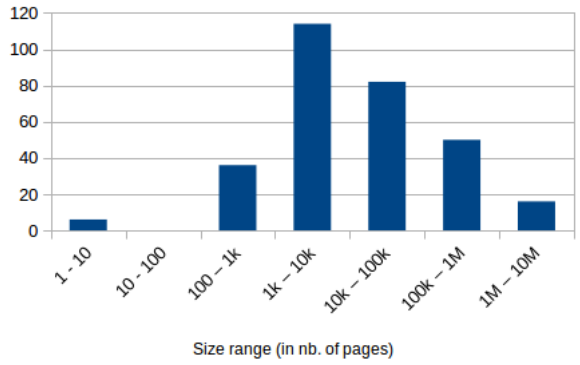

Figure 1: Number of Wikipedias per `size` range (log10 scale).

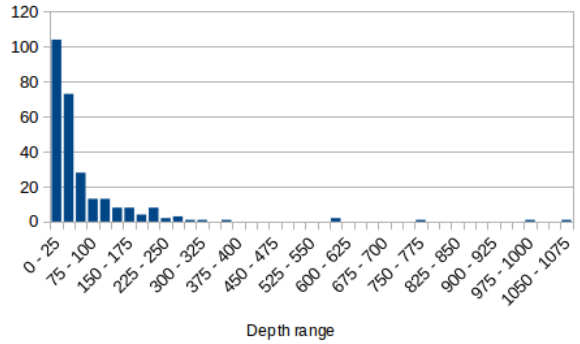

Figure 2: Number of Wikipedias per `depth` range.

pages, redirects etc. It ranges from 0 to 1,063 (Ripuarian Wikipedia). The English Wikipedia has a `depth` of 991. Figures 1 and 2 show the distribution of the Wikipedias according to their size and `depth`.
In fact, we can observe in table 3.1.1.[4] that three of the Top 10 Wikipedias in terms of page number show very poor `depth scores`. This might be explained by the use of translation or bots to produce pages (see, for instance the case of the Swedish Wikipedia (Guldbrandsson, 2013)).
The table also shows that members of small linguistic communities such as the Aragonese or Vepsian ones have seized the opportunity offered by Wikipedia to develop their digital presence.

3.1.2. Identifying the Languages of the Wikipedias

In this section, we present the strategies and issues related to languages and spelling conventions identification in the Wikipedia projects. The examples that follow show that the Wikipedias host both dialectal and scriptural diversities.

Language Tag(s) In its `Language proposal policy`[5], WIKIMEDIA stipulates that each Wikipedia must correspond to a language with a valid ISO 639 1-3[6] code (or, in exceptional cases, a BCP 47 language tag[7] only). What is more :

> "*The language must be sufficiently unique that it could not coexist on a more general wiki. In most cases, this excludes regional dialects and different written forms of the same language.*"

This definition of the accepted languages leads to Wikipedias containing articles written in closely-related dialects. This is the case, for instance, of the "Alemannischi Wikipedia" [8], that contains articles in *Schwyzerdütsch* (Swiss German), *Badisch* (Baden Alemannic), *Elsassisch* (Alsatian dialects), *Schwäbisch* (Swabian German) and *Vorarlbergisch* (Austrian dialect spoken in Vorarlberg). Each article of the Wikipedia is tagged with its corresponding linguistic `category`.
While there exist other examples of multi-dialectal Wikipedias (for instance the Bihari one, which covers more than ten dialects spoken in India and Nepal, or the Occitan one covering a continuum of roman dialects spoken in 4 countries), we have not identified any other Wikipedia in which the articles are explicitly tagged with their corresponding dialects.

Writing Convention(s) The spelling conventions are specific to each Wikipedia. We give here three examples of Wikipedias in which different rules are followed:

- The Alsatian section of the Alemannic Wikipedia contains pages written both in standardized and non-standardized spelling.
- Adversely, the Wikimedia Incubator for Mauritian Creole (ISO 639-3 code `mfe`)[9] displays this note on its front page:

  > "Please use the correct up-to-date standardized spelling of the Mauritian Creole language. Some pages have already been written in as "unstandardized" spelling which need to be replaced."

- The Egyptian Arabic edition (ISO 639-3 code `arz`) is written in Arabic script, yet one page makes the inventory of the articles written in Latin alphabet, intended for "*people who can speak Masry but can only write in the Latin alphabet*"[10].

3.1.3. The Encyclopedic Nature of Wikipedia

As a counterpart for the good quality of the Wikipedias (in terms of the well-formed, grammatical contents they host), contributing to a Wikipedia can be difficult.

[4] The statistics for each Wikipedia are provided by WIKIMEDIA (see `https://meta.wikimedia.org/wiki/List_of_Wikipedias`). The approximate number of speakers per language is the estimation provided by `ethnologue` or was found on the page of the language of the English Wikipedia.

[5] See `https://meta.wikimedia.org/wiki/Language_proposal_policy`.

[6] See `https://iso639-3.sil.org/`.

[7] See `https://tools.ietf.org/html/bcp47`

[8] See: `https://als.wikipedia.org/wiki/Wikipedia:Houptsyte`

[9] See `https://incubator.wikimedia.org/wiki/Wp/mfe/Main_Page`.

[10] See the `Introduction in English` page of the `https://arz.wikipedia.org`.

	Size rank	Size (Nb. articles)	Depth	Approx Nb of native speakers	Active users *
English	1	6,013,707	991	379M	137,409
Cebuano	2	5,378,563	**2**	15M	148
Swedish	3	3,738,252	**7**	10M	2,759
Waray-Waray	11	1,263,914	**4**	2.6M	65
Aragonese	100	36,706	63	10,000	76
Vepsian	167	6,369	39	1,500	23
Hawaiian	195	3,839	8	20,000	14

* "Active Users" are the registered users who have made at least one edit in the last thirty days.

Table 1: Comparison of 7 Wikipedias.

Before all, the Wikipedias are encyclopedias, and the text corpora they represent are a "side effect" of the participation.
The induced expected quality in terms of both content and form, as well as the structured and academic looking environment can represent a barrier for potential contributors.
In fact, encyclopedic articles may not be the most natural content to produce for linguistic communities with recent scriptural tradition.
What is more, as commented by Rémy Gerbert, coordinator of WIKIMEDIA FRANCE, developing Wikipedias for smaller languages faces the obstacle of sourcing the articles, since the sources required to support the article are unlikely to be available in the language of the Wikipedia (personal communication, November 2018).
Finally, it is hard for smaller Wikipedias to cope with the growth of the top ones, hence to be competitive in terms of interest for their users in bilingual contexts. This has for instance been reported in The Digital Language Diversity Project (2017) regarding the preference towards the Italian Wikipedia over the Sardinian one among Sardinian speakers.
There exist interesting initiatives to overcome this issue while taking advantage of existing articles written in a top language. This is for example the case of the experience presented in (Alegria et al., 2013), in which the authors use the Spanish and Basque Wikipedias as corpora, and associate machine translation techniques with human editing performed by volunteers to expand the Wikipedia semi-automatically while creating resources to improve the quality of machine translation.

3.2. The Web as a Corpus

3.2.1. Crawled Corpora for Less-Resourced Languages

One way to address the data bottleneck is to resort to Web crawling. Web crawling for multilingual corpus construction consists in gathering texts from the Web that are further curated and automatically classified by language. For instance, the An Crúbadán project (Scannell, 2007), first initiative of the kind to our knowledge, uses a combination of trigrams, automatically generated lexicons and lists of words specific to a given language, to identify on the Web contents written in 2,228 languages. Goldhahn et al. (2012) combine various techniques, including the bootstrap of corpora through search queries.
Whichever the method chosen to crawl the Web, it is necessary to perform language identification to classify the documents. Both works presented above indeed require statistical information on the distribution of characteristic patterns, such as trigrams, for each language. This kind of information is not always available, especially when it comes to multi-variant languages, composed of similar dialects (eg. the dialects of Occitan) or that can be written with competing orthographies (eg. Cornish).
What is more, the use of crawled corpora is questionable from the legal point of view, as many countries do not recognize the "fair use" applied in the English-speaking world. In order to circumvent the problem[11], some colleagues decided (i) to erase all metadata, and (ii) to scramble the documents[12]. Performing these operations causes a loss of information that results in at least two shortcomings:

- scrambling the documents of a multi-variant language results in building one heterogeneous resource;
- scrambling breaks the structure of the document, hence limiting further use to the sentence level.

Finally, there is no evidence that the best way of processing multi-variant languages is to use an heterogeneous corpus[13].

3.2.2. Social Media-Based Corpora

Social media are a widely used place of expression, hence they can be considered as a valuable source of text corpora. The contents produced on Twitter and Facebook are probably more representative of the conversational use, yet they are not sustainable. As for Facebook, the company does not allow for the free usage of the data and the consent of all the participants should be asked for.[14] Twitter presents different challenges, as Tweets are short texts, which could be considered as quotations and therefore more easily used. However, this does not apply to artistic creations, such as haikus, so the Tweets have to be manually scanned for these. More importantly, to avoid copyright issues the

[11] It is unclear to us to which extent this really solves the legal issue.

[12] See: `https://traces1.inria.fr/oscar/fr/`.

[13] In fact, our own experiments with Alsatian tend to show that training a tool with a small corpus of a given dialect yields to better results on this dialect than using a bigger multi-dialectal corpus.

[14] This has been made clear in a message on the CORPORA list by Eric Ringer, on October 27th, 2015.

Tweets are often referred to by their identifier, but they can be deleted or modified by their creators in the meantime, which generates discrepancies. Besides, the language still needs to be identified, as a user can write tweets in any language they feel is most appropriate to their communication goal.

The availability of linguistic resources being a prerequisite to their further re-usability and longevity, we do not investigate further these sources.

However, initiatives such as the `Nierika` project[15], which was presented at LT4ALL in December 2019, in Paris, could be a solution. This project aims at using social networks to collect linguistic data, while addressing the issue of consent and respecting privacy. As presented by its developer, `Nierika` is "a niche social network on development, which is founded on the objective to collaborate with and support the preservation of all the Mexican indigenous languages". To our knowledge, no result concerning the project has been published so far.

4. Motivations for Crowdsourcing

Crowdsourcing has been successfully used in NLP to compensate the lack of financial means and the unavailability of experts to produce linguistic resources, for example using games with a purpose (Chamberlain et al., 2013) or citizen science platforms, in particular the Language Arc, developed by the Linguistic Data Consorium (LDC)[16]. What is more, it has been repeatedly observed that the success of a crowdsourcing campaign of this kind relies on the openness of the call, that enables to get in touch with few active participants who eventually fulfill the bulk of the chosen task (Chamberlain et al., 2013; Fort et al., 2017; Millour and Fort, 2017).[17] This means that crowdsourcing is not necessarily about finding a way to recruit and motivate a "crowd", and in the context of NLP should not be kept for vast linguistic communities only.

That being said, commonly used microworking platforms such as `Amazon Mechanical Turk` are inadequate for getting in touch with smaller linguistic communities, unlikely to be represented among the microworkers. In fact, there may not exist off-the-shelf solutions to efficiently crowdsource linguistic resources among smaller communities, and such enterprise may lead us to the outer limits of crowdsourcing challenges.

Yet, in a context in which practices are evolving fast and there is probably no expert able to provide sufficient description and resources anyway, involving the speakers in the production of data for their language seems to be the only way out.

In the following, we first present the benefits of involving a variety of speakers to produce linguistic resources, we then detail how crowdsourcing has been successfully used to produce representative data in varied linguistic contexts.

[15]See `https://vaniushar.github.io/about`.

[16]See: `https://languagearc.org/`.

[17]This phenomenon is observed on Wikipedia with for instance around 68K editors on the English Wikipedia, 1K on the Swedish one, 40 on the Cebuano one etc., see `https://stats.wikimedia.org`, figures from December 2019.

4.1. Involving a Variety of Speakers to Produce Linguistic Resources

As multi-variant languages are by definition varied, we believe that the collection process should focus on gathering linguistic productions from a diversity of speakers. This can be observed in works on User Generated Content (UGC), which rely on corpora produced by many speakers to capture the diversity of linguistic practices in a setup where variation with respect to a norm can be observed. We believe a similar approach should be considered for multi-variant languages.

Using crowdsourcing as a way to produce text corpora solves the problem of language identification, since the language is constrained in the first place. Furthermore, the direct contact with participants enables the production of additional metadata such as the dialectal variant or the spelling habit in use. Although in some contexts the speakers may not be able to name the variant in use, they can be asked to point the geographical area on a map. Similarly, when the spelling convention is unknown, we may ask the speaker to indicate their preference towards a suggested spelling over another.

Moreover, crowdsourcing oral languages written by their own speakers allows to avoid the subjectivity of a transcription made by a (field)-linguist, for example. Transcription being an interpretation, we believe that in the context of corpus construction, we prefer having the interpretation of the speakers themselves.

In fact, building resources for endangered languages should focus on developing tools that are actually useful to empower the speakers to use their language.

We believe this cannot be done without collecting data that match today's practice. In fact, developing tools that would work on ancient or literary versions of the language is not what we aim at. Especially, content that may have entered into the public domain because it was published long ago is unlikely to be representative of the current practices. One such example is the corpus for Quechua described in (Monson et al., 2006), which is made of two literary texts first published at the beginning of the 20th century.

Although these corpora are valuable resources, they should not be considered as sufficient. What is more, if we want to be able to involve the speakers in participating into further linguistic processing such as annotation, translation etc., we need to provide them with contents they are comfortable with.

4.2. Crowdsourcing Variation

In this section, we survey how crowdsourcing has been successfully called upon to i) get in touch and involve a variety of speakers to collect data on linguistic variation, ii) collect real world linguistic productions in a controlled setup that matches specific needs and ensures further re-usability of the data.

4.2.1. Oral Data

Crowdsourcing is a common and successful practice when it comes to oral data collection, especially when the goal is to render and document the dialectal variability of a linguistic area. Examples of crowdsourcing of speech corpora

for less-resourced languages include works aiming at collecting the greatest possible variety such as, among others, the work of Cooper et al. (2019) for Welsh dialects (one orthography unifies six dialectal areas).
Such a trend is not surprising, especially considering the present need to document and process languages with a mainly oral tradition. This practice is, to our knowledge, less common when it comes to the collection of written data.
We hypothesize that this might be caused by the most official status taken by the written form over the oral form, even though in practice, spelling in any language is subjected to variations.
Because the transcription time makes the process too costly, and because transcribing crowdsourced oral data is different from crowdsourcing written data produced directly by speakers, we do not investigate further how such technique may be used.

4.2.2. Collaborative Lexicography

The involvement of speakers for collaborative lexicography is a well-studied field, especially when it comes to online dictionaries (Abel and Meyer, 2013).
In fact, there exist numerous projects involving the construction of lexical resources for regional languages and documentation of local variants, based on pre-existing documentation of the dialectal variation. For instance, the *Dictionnaire des mots de base du francoprovençal* uses a standardized supra-dialectal spelling for its entries (Stich et al., 2003).
Following another approach, the "Swiss Italian dialectal Lexicon" [18] has one entry per variant, each of them being linked to a head-term (*capolemma*). Although the designers of this online resource seem to work closely with local speakers, their actual contribution to enriching this resource is unclear (Zoli and Randaccio, 2016).
Duijff et al. (2016) provide feedback on the contribution of speakers for the construction of a Dutch-Frisian dialect dictionary, and especially underline their ability to fill the so-called "lexical-gaps".
These examples show that crowdsourcing can be used to involve a community into collaboratively producing linguistic resources.

5. Building Text Corpora with the Help of the Speakers

Compared to the strategies presented in Section 3., which rely on *collecting* and classifying existing content, we present here strategies developed to actively *build* corpora with the help of speakers.
Crowdsourcing has been used for a variety of tasks as exemplified in Section 4.2., showing that it is possible to involve small linguistic communities into collaboratively producing linguistic data. Yet, to our knowledge, there exists no initiative that aims at producing text corpora for multivariant languages.
In this Section we first present two works of interest with regard to their implicit strategies to collect text corpora.
Then, we present our ongoing work on crowdsourcing linguistic resources and more specifically text corpora for a non-standardized language. After describing the conditions and setup of this experiment, we present the challenges that were encountered as well as an analysis of their potential causes.

5.1. Eliciting Corpora

Crowdsourcing text corpora often resorts to eliciting techniques, such as asking for descriptions to inspire the contributors. In such cases, crowdsourcing can be described as *explicit*, meaning that the goal of the activity is expressed plainly to the participant.
Producing text corpora being a tedious task requiring time and effort, Niculae and Danescu-Niculescu-Mizil (2016) and Prys et al. (2016) have come up with original ideas to crowdource text corpora *implicitly*. The first article presents `Street Crowd`, an online game which objective is to identify the location where a picture was taken. This search towards the correct location is done collaboratively, with multiple participants giving their opinion and possibly debating the solution. The crowdsourced corpus is here composed of the conversations between the participants. The second article presents an online spell and grammar checker for Welsh, used as such by speakers. The corpus collected here is the input to be spellchecked. This strategy appears as particularly efficient to collect diverse data in terms both of form and content.

5.2. Crowdsourcing Cooking Recipes

We have focused in previous work on producing corpora collaboratively annotated with part-of-speech for under-resourced languages (Millour and Fort, 2018; Millour and Fort, 2019). Our experiments involved Alsatian, a continuum of Alemannic dialects spoken in Alsace, a diglossic French region, and Mauritian Creole, a French-based Creole spoken mostly in Mauritius. A flexible spelling system called Orthal has been developed for Alsatian (Crévenat-Werner and Zeidler, 2008) and a standardized spelling (*Lortograf Kreol Morisien*) is promoted by the Mauritian Creole Academy (*Akademi Kreol Morisien*) (Police-Michel et al., 2012) and supported by the Mauritian government. Although, to our knowledge, there exists no precise statistics on the use of these spelling recommendations, neither of them seem to be widespread among the Alsatian and Mauritian Creole speaking communities (Saarinen, 2016; Erhart, 2018). This lack of standardization translates into the coexistence of alternative spellings for many words, expressing both the dialectal and scriptural variations at stake.
For the sake of sustainability, we chose to provide the speakers with text corpora that was distributed under a clear license so that we would be able to share its annotated version.
Yet, in both cases, we were rapidly limited by the small size of the available corpora. Additionally, these annotating experiments confronted us with two issues:

1. The discomfort expressed by participants: some of them struggled annotating sentences that were not written accordingly to their own practice of the language, either in terms of dialect or spelling habit.

[18] *Lessico dialettale della Svizzera Italiana*, see `http://lsi.ti-edu.ch/lsi/`.

2. The unbalance in variants in our corpora, extracted from the Alemannic Wikipedia. The taggers trained on the crowdsourced annotated corpus were biased towards the over-represented variant (Millour and Fort, 2018).

This brought us to crowdsource additional text corpora. We chose to collect cooking recipes to elicit production. We found three benefits in involving the speakers into collaborative text corpus creation. First, text collection would naturally increase the size of the available corpora. Second, since the participants would annotate their own texts, they would not feel the discomfort expressed above. Third, involving speakers of various linguistic profiles would increase the representativeness of our corpus.

Along with the corpus collection, we added a feature called "I would have said it like that" which enabled the participants to suggest an alternative spelling for any word present on the crowdsourced corpus. This feature is exemplified in figure 3.

The corpus collection experiment did not yield the expected results, since less than 10 participants entered recipes. In fact, our first experiment with crowdsourcing, which was about annotating existing corpora with the universal part-of-speech tagset, was more successful than our second attempt, with more than 50 participants producing up to 19,000 annotations (Millour and Fort, 2018)

The hypothesis we had made that a "non-linguistic" task would be more attractive than an annotation task was not confirmed by our experiments, even though our second platform was designed with more attention, was publicized in the local newspaper and blogs, hence benefiting from better advertising. Actually, the advertising made on our second platform brought additional participants to the annotation task.

Interestingly, the feature aiming at collecting spelling alternatives on pre-existing words received more interest and 367 alternative spellings were provided for 148 words (Millour and Fort, 2019).

Overall, our experience in crowdsourcing linguistic material for Alsatian leads us to suspect that producing text corpora might be harder a task than we thought, and requires more careful design.

5.3. Why are the Speakers Reluctant to Participate?

To understand the unequal participation observed on our crowdsourcing platforms, we conducted online surveys.

We were inspired by the `Digital Language Diversity Project` (DLDP), which, with the support[19] of WIKIMEDIA FRANCE, has conducted four surveys to understand how the digital presence of four "minority languages" could be developed. The languages which received attention were Breton (200 replies), Basque (428 replies), Karelian (156 replies), and Sardinian (596 replies) (Soria et al., 2018).

From our part, we have conducted, in parallel with the crowdsourcing experiments, two surveys to get a better insight on how the Alsatian and Mauritian Creole speaking communities felt about the use of their language online. A great majority of the members of both linguistic communities are at least bilingual with a language that is taught in school and standardized (like French or English).

To enable comparison with the surveys created by the DLDP, we kept most of their structure, to which we added a focus on:

- the relationship of the speakers with the written form of their language,
- their perception of dialectal and scriptural variety,
- their knowledge and use of the existing spelling standards.

The first survey, entitled "Alsatian, the Internet and you" [20] received 1,200 replies. The second is entitled "Mauritian Creole and its digital presence"[21] and received 144 replies. Both surveys were published in French, the Alsatian community (Huck et al., 2007) and 98% of the Mauritius population (Atchia-Emmerich, 2005) being bilingual with French.

Most of the respondents of the surveys led by DLDP are language activists, professionally involved with their language: 66% for Breton, 65.3% for Basque, 60.7% for Sardinian, 48.7% for Karelian (for which 69.9% of the respondents state they take part in either a revitalization or protection activity related to Karelian). As highlighted by the authors of these studies, this might introduce a strong bias.[22]

In comparison, 25% of the Alsatian respondents and 18% of the Mauritian respondents to our surveys state they have either a professional or associative involvement with their language.

Note that there is no widely spread spelling standard for Alsatian, Mauritian Creole, Karelian and Sardinian, while there exist consensual orthographies for Breton (the Peurunvan orthography) and for Basque (*euskera batúa*, literally the "unified basque").

Interestingly, the survey led on Basque and Breton shows no difference between spoken and written self-evaluation. As for Mauritian Creole, 38% of the respondents had never heard of the spelling conventions defended by the Mauritian Creole Academy. Regarding Alsatian, 69% of the respondents claim they had never heard of the Orthal spelling system. Of the 73% who evaluate their oral proficiency as good, only 33% also evaluate their writing proficiency as good, while 49% evaluate it as medium, and 17% as weak.

[19] See `https://www.wikimedia.fr/2016/08/03/digital-language-diversity-project-et-wikimedia-france/`.

[20] In French "*L'alsacien, Internet et vous*", available here: `https://framaforms.org/sondage-pratiques-linguistiques-en-ligne-1546808704`

[21] In French "*Le créole mauricien et sa présence en ligne*", available here: `https://framaforms.org/sondage-le-creole-mauricien-et-sa-presence-en-ligne-1555054850`.

[22] "Language activists tend to be intentionally more assertive in their use of the language and, as a consequence, they can't represent average speakers." (Soria et al., 2018)

Figure 3: Spelling addition (1) and visualization (2) on a crowdsourced recipe in Alsatian (highlighted words present at least one additional variant).

It therefore seems that the average speakers –not the language activists– under-evaluate their ability to write their own language and might be reluctant to write it on a platform developed by researchers.

Depending on the strategy chosen to crowdsource (the task to perform can either be explicit, or hidden under another purpose, hence implicit), designers should bring an extra care to raising awareness about the urge to develop linguistic resources.

They also should make a pedagogical effort to convince the speakers that the way they write their language cannot be wrong and that we need their input to develop systems dealing with the language as it is used today.

6. Conclusions and Perspectives

Oral languages are more and more written by their speakers, especially on digital media. This is an opportunity for us, as researchers in linguistics and NLP, both in terms of needed applications (eg. word prediction) and collection of language resources.

In this context, and since very little research of this kind has been carried out, it is still unclear whether crowdsourcing to encourage data production is worth the effort. On the other hand, we have seen that the material spontaneously produced by the speakers and made available online is often insufficient to fulfill the NLP researchers needs, especially in the context of non-standardized languages. In fact, we believe that initiatives involving speakers are more likely to produce usable material.

During our experience with crowdsourcing, we have experimented that speakers seem to be reluctant to provide us with language data, as they feel like they do not know how to write their language properly. These psychological barriers should be addressed by researchers in order to overcome the lack of diversity in the freely available data we need.

A solution to this is to use a real game as support for crowdsourcing, so that the speakers "forget" that they are participating to a research experiment. We thus developed a prototype of a role-playing game (RPG), which aim is both to foster the inter-generational transmission of non-standardized languages and to collect lexicon (including multi-word expressions) and variants for the language (Millour et al., 2019). This game will be made freely available for translation and use in any language, so that, hopefully, kids will be proud to speak and write their family language.

7. Bibliographical References

Abel, A. and Meyer, C. M. (2013). The dynamics outside the paper: user contributions to online dictionaries. In *Proceedings of the 3rd eLex conference 'Electronic lexicography in the 21st century: thinking outside the paper*, pages 179–194.

Alegria, I., Cabezon, U., Fernández de Betoño, U., Labaka, G., Mayor, A., Sarasola, K., and Zubiaga, A., (2013). *Reciprocal Enrichment Between Basque Wikipedia and Machine Translation*, pages 101–118. 02.

Atchia-Emmerich, B. (2005). *La situation linguistique de l'île Maurice: Les developpements récents à la lumière d'une enquête empirique*. Ph.D. thesis, Universität Erlangen-Nürnberg, 03.

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013). Using games to create language resources: Successes and limitations of the approach. In Iryna Gurevych et al., editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.

Cooper, S., Jones, D. B., and Prys, D. (2019). Crowdsourcing the paldaruo speech corpus of welsh for speech technology. *Information*, 10(8):247.

Crévenat-Werner, D. and Zeidler, E. (2008). *Orthographe alsacienne - Bien écrire l'alsacien de Wissembourg à Ferrette*. Jérôme Do Bentzinger.

Duijff, P., van der Kuip, F., Sijens, H., and Visser, W. (2016). User contributions in the online dutch-frisian dictionary. In *Proceedings of European Network of e-Lexicography (Enel) COST Action (WG1 meeting)*, Barcelona, Spain.

Erhart, P. (2018). Les émissions en dialecte de france 3 alsace : des programmes hors normes pour des parlers hors normes ? In *Les Cahiers du GEPE*. Strasbourg : Presses universitaires de Strasbourg.

Fauzi, A. I. and Puspitorini, D. (2018). Dialect and identity: A case study of javanese use in WhatsApp and line. *IOP Conference Series: Earth and Environmental Science*, 175:012111, jul.

Fort, K., Guillaume, B., and Lefèbvre, N. (2017). Who

wants to play Zombie? A survey of the players on ZOMBILINGO. In *Games4NLP 2017 - Using Games and Gamification for Natural Language Processing*, Symposium Games4NLP, page 2, Valencia, Spain, April.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings ofthe 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 759—-765, Istanbul, Turkey. European Language Resources Association (ELRA).

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Guldbrandsson, L. (2013). Swedish wikipedia surpasses 1 million articles with aid of article creation bot. *Wikimedia blog*, 17.

Huck, D., Bothorel-Witz, A., and Geiger-Jaillet, A. (2007). *L'Alsace et ses langues. Éléments de description d'une situation sociolinguistique en zone frontalière*. Université de Strasbourg.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only.

Lillehaugen, B. D., (2016). *Why write in a language that (almost) no one can read? Twitter and the development of written literature*, volume 10, pages 356–393. University of Hawaii Press.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Millour, A. and Fort, K. (2017). Why do we Need Games? Analysis of the Participation on a Crowdsourcing Annotation Platform. In *Games4NLP*, Valencia, Spain, April.

Millour, A. and Fort, K. (2018). Toward a Lightweight Solution for Less-resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing. In *Proceedings of11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Millour, A. and Fort, K. (2019). À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. In *Revue TAL : numéro spécial sur les langues peu dotées*, volume 59-3. Association pour le Traitement Automatique des Langues.

Millour, A., Grace Araneta, M., Lazić Konjik, I., Raffone, A., Pilatte, Y.-A., and Fort, K. (2019). Katana and Grand Guru: a Game of the Lost Words (DEMO). In *9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'19)*, Poznań, Poland, May.

Monson, C., Llitjós, A. F., Aranovich, R., Levin, L. M., Brown, R., Peterson, E., Carbonell, J. G., and Lavie, A. (2006). Building nlp systems for two resource-scarce indigenous languages : Mapudungun and quechua. In *Proceedings of5th SALTMIL Workshop on Minority Languages*, pages 15–24, Genoa, Italy.

Niculae, V. and Danescu-Niculescu-Mizil, C. (2016). Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'2016)*, pages 568–578, San Diego (CA), USA, June. Association for Computational Linguistics.

Outinoff, M. (2012). English won't be the internet's lingua franca. In *Towards the Multilingual Cyberspace*, pages 171–178. Vannini, Laurent and Le Crosnier, Hervé, c&f éditio edition.

Police-Michel, D., Carpooran, A., and Florigny, G. (2012). *Gramer Kreol Morisien*. Akademi Kreol Morisien, Ministry of Education and Human Resources.

Prys, D., Prys, G., and Jones, D. B. (2016). Cysill arlein: A corpus of written contemporary welsh compiled from an on-line spelling and grammar checker. In *Proceedings ofthe 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 3261–3264. Portorož, Slovenia, European Language Resources Association (ELRA), May.

Rivron, V. (2012). L'usage de Facebook chez les Éton du Cameroun. In *Net.lang Réussir le cyberespace multilingue*, pages 171–178. Vannini, Laurent and Le Crosnier, Hervé, c&f edition.

Saarinen, R. (2016). La distribution des fonctions des langues dans un contexte multilingue : cas de l'Île maurice. Master's thesis, Faculté des Lettres, Université de Turku, Turku, Finland.

Scannell, K. P. (2007). The crúbadán project: Corpus building for under-resourced languages. In *Proceedings of the 3rd Web as Corpus Workshop: Building and Exploring Web Corpora*, volume 4, pages 5–15, Louvain-la-Neuve, Belgium, September.

Shekhar, S., Sharma, D. K., and Beg, M. S. (2018). Hindi roman linguistic framework for retrieving transliteration variants using bootstrapping. *Procedia Computer Science*, 125:59–67.

Soria, C., Quochi, V., and Russo, I. (2018). The DLDP survey on digital use and usability of EU regional and minority languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Stich, D., Gouvert, X., and Favre, A. (2003). *Dictionnaire des mots de base du francoprovençal : orthographe ORB supradialectale standardisée*. Le Carré.

The Digital Language Diversity Project. (2017). Sardinian — a digital language? In *Reports on Digital Language Diversity in Europe*. Editors: Claudia Soria, Irene Russo, Valeria Quoch.

Tobaili, T., Fernandez, M., Alani, H., Sharafeddine, S., Hajj, H., and Glavas, G. (2019). Senzi: A sentiment analysis lexicon for the latinised arabic (arabizi). in: International conference recent advances. In *Proceedings ofRecent Advances in Natural Language Processing (RANLP'2019)*, Varna, Bulgaria, September.

Tournadre, N. (2014). The tibetic languages and their classification. In *Trans-Himalayan linguistics: Historical and descriptive linguistics of the Himalayan area*. Owen-

Smith, Thomas / Hill, Nathan.

van Esch, D., Sarbar, E., Lucassen, T., O'Brien, J., Breiner, T., Prasad, M., Crew, E., Nguyen, C., and Beaufays, F. (2019). Writing across the world's languages: Deep internationalization for gboard, the google keyboard. *arXiv preprint arXiv:1912.01218*.

Zoli, C. and Randaccio, S. (2016). The context of use of e-dictionaries for the minority languages of italy (case study). In *Proceedings of European Network of e-Lexicography (Enel) COST Action (WG3 meeting)*, Barcelona, Spain, March.