Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 106–110
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Poio Text Prediction: Lessons on the Development and Sustainability of LTs for Endangered Languages

Vera Ferreira, Pedro Manha, Gema Zamora
Interdisciplinary Centre for Social and Language Documentation (CIDLeS)
Rua do Remexido, Loja 15, 2395-174 Minde, Portugal
vferreira@cidles.eu, pmanha@cidles.eu, gzamora@cidles.eu

Abstract

2019, the International Year of Indigenous Languages (IYIL), marked a crucial milestone for a diverse community united by a strong sense of urgency. In this presentation, we evaluate the impact of IYIL's outcomes in the development of LTs for endangered languages. We give a brief description of the field of Language Documentation, whose experts have led the research and data collection efforts surrounding endangered languages for the past 30 years. We introduce the work of the Interdisciplinary Centre for Social and Language Documentation and we look at Poio as an example of an LT developed specifically with speakers of endangered languages in mind. This example illustrates how the deeper systemic causes of language endangerment are reflected in the development of LTs. Additionally, we share some of the strategic decisions that have led the development of this project. Finally, we advocate the importance of bridging the divide between research and activism, pushing for the inclusion of threatened languages in the world of LTs, and doing so in close collaboration with the speaker community.

Keywords : Less-Resourced/Endangered Languages

1. Motivation

2019, the International Year of Indigenous Languages (henceforth IYIL) marked a crucial milestone for everyone involved in the cause of endangered languages, either as researchers, teachers, or activists. We are a diverse community of experts and implementers united by a strong sense of urgency and, although our immediate interests do not always coincide, we have gathered momentum and we must make the most out of it.

There are around 7000 languages spoken today but, at the current rate, it is estimated that half of them will vanish in the next one or two generations. UNESCO has taken on, for good or for bad, the responsibility of following up on these statistics and map the languages of the world to raise awareness around the issue of language endangerment and its impact on minority groups and their environment. Following this line of action that started in the 1990s, UNESCO and many partnering institutions celebrated language diversity in 2019 and called for the empowering of indigenous peoples through the strengthening of their languages. In November 2019, UNESCO released a strategic outcome document summarising conclusions and recommendations drawn from consultations carried out during the IYIL. The ambitious and optimistic tone of this document is indeed refreshing, especially since we are used to the media treating the issue of endangered languages with a rather melancholic contentment. Moreover, the document reaches to a wide variety of stakeholders and it addresses technology developers directly. In its conclusion V, UNESCO has called for LT developers to "develop advanced tools for collection and analysis of language data as well as for the transliteration and annotation of multi-modal content", "supply necessary tools for advanced translation", and "extend and refine current language technologies as well as designing new ones, and developing necessary algorithms, applications and systems to support indigenous peoples in their own use of the internet and social media networks" (UNESCO, 2019 pp16-17).

Considering these recent developments, we present our experiences with the development of LTs for endangered languages with special attention to the Poio project. The aim of this paper is to encourage participation among endangered language experts that are not familiar with LTs, and to make explicit these technologies' potential for impact and innovation.

In section 2, we give a brief overview of the motivations and concerns of the field of Language Documentation and introduce the work of our institution. Section 3 describes Poio, one of our long-standing projects, and some aspects of its roadmap. Section 4 to 6 describe the challenges that the institution has faced regarding the development of LTs for endangered languages and provide examples of our team's strategic approach. Finally, we conclude the presentation by summarising the lessons we have learnt during the development of Poio and other tools, and giving the message that it is possible to create LTs for speakers of endangered languages that are sustainable in the long-term.

2. CIDLeS' Background and Roadmap

The Interdisciplinary Centre for Social and Language Documentation (CIDLeS) is a non-profit institution founded in January 2010 in Minde (Portugal) by a group of national and international researchers. From the moment of its foundation, CIDLeS aimed at improving and deepening research in two linguistic areas: language documentation and linguistic typology. Besides the documentation, study and dissemination of European endangered and minority languages, CIDLeS (CIDLeS Media Lab only until recently) is also engaged in the development of language technologies for scientific and didactic work on lesser-used languages.

Language Documentation was recognized and established as a linguistic discipline in the late 90s of the 20th century. However, the areas of interest as well as its subjects of study (e.g. description and classification of linguistic features from around the world) have been of interest to all linguists, especially to those who worked in the area of typology or anthropological lingusitics with a broad experience of fieldwork. It is in this context that CIDLeS was founded, and its work stands out for the application of language documentation methods to European languages (Minderico, A Fala or Bavarian are some examples), which

tend to be overlooked in a discipline that draws many of its methods from anthropology and ethnography.
CIDLeS also stands out for its push for community-driven maintenance, and its investment in LTs. It is widely recognised that language documentation and language maintenance/revitalisation efforts could and should work in tandem. However, some practicalities such as funding availability or workload make this synergy more complex than it seems. While most linguists recognise their ethical responsibility towards the communities from which they collect data, they often lack the means to provide said community with resources that can help keep the language vital (Leonard, 2018). On the other hand, language revitalisation experts and activists might sometimes overlook the potential of documentation materials as resources for the speakers due to their theoretical inaccessibility. CIDLeS tries to bridge that divide by developing software for speakers of lesser-used languages that can re-use data originally collected for linguistic research.
We believe that bridging this gap ties closely with conclusion and goal V in UNESCO's recommendations as it "[…] allow[s] the development of technologies specifically adapted to the characteristics of indigenous languages, which in turn will strengthen and underpin the status of these languages" (UNESCO, 2019 pp16).

3. Poio API, Poio Corpus and Poio Text Prediction

We do not want to limit ourselves to developing the necessary LT tools for local and minor languages and language varieties as mere aids for communication. Our goal is to use successful technology to teach, revitalize and therefore boost the use of minority languages. People should not only be able to communicate in their natural, native tongue, but technology should also assist the renewal of local languages and cultures by allowing people to actively teach, learn, extend and spread their language in their community (Ferreira, 2016). We see language diversity and multilingualism as one building block to empower local communities and their cultural identity and thus realize their cultural and economic potential in a globalized world.
Poio is the name of a project, under the responsibility of Peter Bouda, with several open source subprojects which develop LTs (Bouda, Ferreira, and Lopes, 2012). Our aim is to give people the ability to use their mother tongue in everyday, electronic communication in the digital world, no matter where they are and whatever language they speak. Poio provides the technological basis to process language data from a wide range of sources (e.g. language documentation corpora, Wikipedia, retro-digitized and digital dictionaries, etc.) for applications and research workflows. This includes, for example, the possibility of extracting data from ELAN transcription files, which are widely used for transcribing language documentation recordings and creating multimodal corpora.
Poio consists of several subprojects that make possible to process and manage language data, to extract corpora from diverse data sources and to calculate language models for the online tools. At the basis of the Poio project are our two scientific Python packages, Poio API[1] and Graf-python, which allow us to manage data from a wide range of sources (eg. ELAN, Toolbox. TypeCraft XML) and convert them into GrAF for interoperability and further analysis. An example of an straightforward application is the conversion from Toolbox to TypeCraft with the aid of a JSON mapping file. Once the annotations are in TypeCraft, it is possible to share and further annotate the data in a group and/or to create web-based applications.
The Poio Corpus is a collection of data in under-resourced languages extracted from Wikipedia, websites, and dictionaries. It is available for free download in ISO format. With this corpus at its foundation, the Pressagio library, also in Python, predicts text based on n-gram models.
The data management and text prediction functionalities of Poio API and Pressagio are available as the web service Poio Web API. Poio Text Prediction is its equivalent for end users, which can be accessed online[2]. This is nowadays the visible face of the Poio project and the arm of the project that we are presenting here as an example of an LT devised with the needs of speakers of endangered languages in mind.
The text prediction that Poio offers can be easily used on the desktop and on mobile phones and tablets. Users can write their texts by clicking on the offered prediction and copying the texts to their email editor or messaging app. Figures 1 and 2 show how the predictions are displayed differently on computers or mobile devices. When accessing the service on a computer, the user can select the appropriate prediction by clicking with their mouse or using the F keys. On mobile devices, however, the user taps on the prediction the same way they would do with their usual predictor in Portuguese.

Figure 1: Example of Poio Text Prediction in Minderico.

[1] Poio API's development was part of the curation project F-AG3 within CLARIN-D. The latest version is available for download at https://github.com/cidles/poio-api and the documentation with use examples is available at https://poio-api.readthedocs.io/en/latest/introduction.html .
[2] Service available at https://www.poio.eu/. Documentation available at https://poio.readthedocs.io

Figure 2: Example of Poio in a mobile device.

The need to communicate seamlessly on social media or via texts and to do so in the language of one's choice goes often beyond adding special characters or using symbols. The traditional predicting keyboards support major languages like English, Spanish, or Chinese. Thus, the users that belong to lesser-used and under-resourced language speaking communities are not able to use their native tongue in an easy and successful way. Furthermore, these keyboards may work against the user's own effort to remember a word in their language by giving a suggestion in the national language or "correcting" its spelling.

Currently, manufacturers of mobile devices and operating system owners are gradually opening their systems and devices to developers. This tendency is becoming more evident for instance in the domain of virtual keyboards used in current mobile operating systems, with the possibility of creating third party keyboards for use alongside the default one. Additionally, open-source or more customisable alternatives are increasingly available on the market. These factors make clear to us that the development of Poio Text Prediction and its enhancement for mobile devices have now more potential than ever.

For all of us who take the English or Portuguese (or any language of major communication) predictive text engines in our phones for granted, this might seem a minor change. However, it would have a great impact among all the users who do not feel confident enough to write in their own minority language, or that simply do not know they can do it. Predictive text is also practical for learners and semi-speakers of the language, because it can help finish a sentence or remember a word without having to force oneself to look it up on a dictionary or switch to another language.

Our next steps within the Poio project are 1) improving the text prediction system for the languages already supported, 2) increasing the number of languages supported, and 3) working on the development of an offline service stored on mobile devices and desktops to allow text entry in under-resourced languages in different technological contexts. (To be able to use Poio, the user needs an active internet connection – one of the issues CIDLeS team is working on at the moment.)

At the moment of writing, Poio Writetyper is available in Afrikaans, Aragonese, Asturian, Basque, Bavarian, Chechen, Corsican, Ewe, Faroese, Friulian, Haitian, Irish, Ligurian, Lombard, Low German, Lower Sorbian, Luxembourgish, Manx, Minderico, Norther Frisian, Romansch, Saterfriesisch, Scottish Gaelic, Upper Sorbian, Venetian, Welsh, and Western Frisian.

There is a steady interest in Poio that has allowed the team to further its development for several years now, but sustainability is indeed the Achilles heel for many technology projects aimed at endangered languages or developed around one single community. In the following sections, we elaborate on the three main challenges that the team has faced since the early stages of development and some strategic decisions that we have decided to carry on to our future endeavours. We hope that CIDLeS and Poio's example can encourage more groups of experts to invest time in data mobilisation and that it can help others identify opportunities for interdisciplinary collaboration and community involvement.

4. Challenge I: Do LTs Have to Take a Back Seat to Research?

Since its conception in 2013, the development of Poio and other CIDLeS projects have encountered a series of challenges that we believe are common to most LTs for speakers of endangered languages. The decreasing number of users or insufficient literacy in some communities might seem the most apparent problems. However, these respond to deeper political and developmental problems and, in today's globalised economy, any innovative practice aimed at preventing loss of diversity faces similar challenges. Here, we would like to assess those issues that we developers and implementers face in the specific case of LTs for endangered languages.

The first challenge is the gap between current research trends in Linguistics and what the speaker community actually finds useful. While there is a growing interest in academia for automating certain aspects of the language documentation process and enhancing the analysis of endangered languages with digital methods (Michaud et al., 2018), the outputs of these projects (e.g. grammars, corpora, highly technical dictionaries, etc.) do not always benefit the speaker community or respond to their needs (Leonard, 2018). On the other hand, programmers and computer scientists working with linguists are able to produce sophisticated databases, mine data and aid the methodology of a given research project, but are rarely involved in the ethically-charged process of giving back to the community. This is not to say that academics are not doing enough, but that they cannot be expected to do everything.

In this scenario, LTs for speakers of endangered languages have little time and space for recognition and development as they are not necessarily useful to answer a research question. The Poio project itself was initially motivated by the need to find new methods to mine data for typological and morphological research. However, Poio Text Prediction has proved advantageous for its developers and the CIDLeS community in ways that we did not necessarily predict. First, being so easy to deploy and test makes it a

good instrument for networking and a great conversation starter. As such, it gives visibility to its developers' work beyond the academic sphere. And most importantly, it has a tangible impact in the community in the form of digital presence. This is especially true for communities that have little to no representation online as they lack the computer literacy or the networks to foster the growth of their languages in places like Wikipedia or Twitter.

For example, because it is possible to use Poio API to process language corpora or archival collections, the transformation of these materials into a text prediction service makes these resources accessible to a wider public. Furthermore, it does it in a way that is coherent with the community's context and it has the potential to appeal the younger generations. The basic requirements to achieve this are: 1) the source material must cover a fairly varied range of topics collected from speakers of different ages and genders, 2) the source texts must be anonymised, and 3) the potential user community must already have an interest in writing in their language, and 4) the potential user community has steady access to the internet in order to use the service.

This is also an example of how the development of an LT can derivee naturally from language documentation efforts, making the most out of the funding available, easing the relationship between linguists and community members, and producing content that is automatically tailored to its users. Given the sense of urgency that currently dominates the fields of language and intangible cultural heritage documentation, we believe that finding ways to bridge data collection, analysis, and community engagement as seamlessly as possible should be a priority for developers and implementers. In the following section, we elaborate further on the issue of community engagement.

5. Challenge II : Are LTs the New « Holy Grail » of Language Maintenance ?

The second challenge is the political pressures that surround endangered languages and the often counterproductive belief that the way to maintain the language is making it official and institutionalise its learning. Schools, and in particular those located in rural areas or that serve marginalised communities, have very limited resources and are a biased battleground for the endangered language to compete against the national language or English. On the surface, teaching the language in school is a sign of prestige, but it does not guarantee the natural transmission of the language. In a similar way, LTs play a crucial role in giving a sense of prestige to the language, especially among the younger generations who are constantly exposed to technologies that compete for their time and attention. Games, online resources, and electronic teaching aids have gained prominence since the early 2000s (Eisenlohr, 2004). However, just like we know that school education alone does not necessarily guarantee the maintenance of an endangered language, we must be cautious not to attribute to technology the capability of keeping a language vital just on its own. LTs must be part of a cohesive effort for improving the social status of a language and foster its use.

Nowadays, given the highly competitive market that we live in, the social status of a given language (endangered or not) is often based on the answers you can give to questions such us "Which benefits can I access through this language?", "Can this language get me better employment opportunities?", "How does learning/speaking this language make me look in front of my friends?". While not even the most appealing, innovative LT can give positive answers to these questions on its own, at CIDLeS we are learning to uncover the skills used throughout the stages of data collection, corpus building, or LT development and re-package them in ways that may appeal to more members of the community. This approach was inspired by the young Minderico speakers and learners that interned at CIDLeS in its early days and for whom their experience working in the development of Poio has been an asset in their careers.

The practice of involving the speakers actively in a linguistic research project is not new; see Harvey (2019) for a recent example. Nevertheless, we believe that there is still a lot of work to be done even in Western and/or urban contexts. Coders and developers could find here a platform to become mentors and strengthen ties with the communities they serve.

The success of an LT for endangered languages will always depend on grassroots, socially oriented groundwork and the motivation of the speaker and learner community. On top of this, the development of such technology has to be culturally contextualised. Analysing and taking into account the speakers motivations and networks, not only at the beginning and end of the project as potential users, but throughout the development process as stakeholders is a strategic decision. While we understand that this decision might present other logistics challenges, especially when introducing new technologies, it lays the groundwork for fruitful collaborations long-term.

6. Challenge III: Funding is in an Uncomfortable Grey Area

Academics in the area are already working at their maximum capacity and, despite the best intended advocacy efforts of the international community, the attitudes towards endangered languages are still pessimistic. These are just two of the factors that contribute to the biggest hurdle in the development of LTs for endangered languages: funding for community-oriented projects is very limited and it rarely takes into account the long-term sustainability of the outcomes. Furthermore, the private sector sees little to no gain in supporting these initiatives as the general attitude towards endangered and minority languages is that they do not have marketable value.

As we mentioned in the introduction to this paper, UNESCO has made a rather ambitious call for researchers and developers to work on sophisticated LTs that support indigenous peoples on the use of their native language. They also recommend all stakeholders to "encourage collaboration between indigenous people, researchers, and industry" and to "make it an urgent priority to encourage the donor community, intergovernmental organizations, and other stakeholders towards mobilizing additional financial resources and establishing new funding mechanisms and incentives for activities and projects on indigenous language issues" (UNESCO, 2019 pp.18-19).

We believe that this collaboration is highly beneficial for all parties and that it will keep growing over time. However, whether the funding will be more evenly distributed, or if specially designated funding will be allocated is something that we have yet to see. In the meantime, and in case that the funding situation does not

improve significantly, we are exploring two options: crowdfunding and funding opportunities with a focus on social entrepreneurship.

Small contributions in the form of annual or monthly subscriptions are a way of maintaining the Poio Text Prediction service online. The project responsible is currently piloting different tiers with a focus on expanding the languages available and the quality of the service.

Simultaneously, CIDLeS is exploring whether Poio and other LTs for under-resourced languages could fit in and be benefited from funding pools aimed at community development, compulsory education, and further education. This way, we seek to make the most out of the working relationships we have stablished with stakeholders from communities outside Minde and to design an LT project that has endangered language speakers at its core.

7. Conclusion

Throughout this paper we have outlined the challenges commonly faced when developing LTs for endangered languages. However, we wanted to present Poio as a success story, not only because we have been working with it for 7 years, but because its scope is clearly in line with UNESCO's recommendations. UNESCO's white paper is in this case an assertion of what we and many other experts have been doing right so far and, here, we have offered an overview of the strategic decisions that we have taken in the development of one of our most successful projects.

We are optimistic that this new push for recognition will mean a positive stimulus for LTs for under-resourced languages and that they will encourage a new generation of developers to take an interest in supporting endangered languages research and maintenance with their work. Experience has taught us that, as developers, we cannot ignore the factors that make endangered languages endangered, and that resources are way too limited to risk investing time in a language without proactively involving its speakers throughout the development process.

We have seen that, in the case of endangered languages, LTs fit almost awkwardly between research and community development, with no interest from the private sector. However, these LTs are far from isolated projects developed around single small languages. Instead, they are part of a larger effort and have the potential for big societal impact. While our main objective is to make solid technologies, we are also project managers and advocates, and the future success of these technologies depend on acknowledging and exploiting our role within the communities we serve.

If we want LTs for endangered languages to be successful and sustainable, we must continue our work with linguists in order to keep the quality of the documentation and data collection at the highest standards. Also, equally importantly, we must make the most out of these resources and use them to create technologies that empower their users to assert themselves in the language of their choice. As long as there is grassroots interest, there will always be room for endangered languages in LT development.

8. Bibliographical References

Bouda, P., Ferreira, V. and Lopes, A (2012). Poio API - An annotation framework to bridge Language Documentation and Natural Language Processing. In F. Mambrini, M. Passarotti and C. Sporleder (eds.). Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities. Lisboa: Edições Colibri, 15-26

Eisenlohr, P. (2004). Language Revitalization and New Technologies. *Cultures of Electronic Mediation and the Refiguring of Communities. Annual Review of Anthropology.* 33. 21-45.

Ferreira, V . (2016). The importance of new technologies in the revitalization of Minderico. In J. Olko, T. Wicherkiewicz & R. Borges (eds.). *Integral Strategies for Language Revitalization*. Warsaw: University of Warsaw, pp565-580.

Harvey, A. (2019). Gorwaa (Tanzania) —Language Contexts. In P. K. Austin (ed.) *Language Documentation and Description, vol 16*. London: EL Publishing, pp127-168

Leonard, W. Y. (2018). Reflections on (de)colonialism in language documentation. In B. McDonnell, A. L. Berez-Kroeker, & G. Holton. (Eds.) *Reflections on Language Documentation 20 Years after Himmelmann 1998*. Language Documentation & Conservation Special Publication no.15. Honolulu: University of Hawai'i Press. pp55-65.

Michaud, A., Adams, O., Cohn T.A., Neubig, G., and Guillaume, S. (2018). Integrating Automatic Transcription into the Language Documentation Workflow: Experiments with Na Data and the Persephone Toolkit. *Language Documentation & Conservation* Vol. 12 pp393-429

UNESCO (2019). *Strategic Outcome Document of the 2019 International Year of Indigenous Languages*. Annex to General Conference 40th session. Paris.