Yezh Ar Vro -The language of the country: Building the appropriation of data collection applications *

Mélanie Jouitteau, Jean-Yves Antoine, Loïc Grobol, Alice Millour

Yezh Ar Vro - The language of the country: Building the appropriation of data collection applications*

Mélanie Jouitteau[1] Alice Millour[2] Jean-Yves Antoine[3] Loïc Grobol[5]
(1) CNRS, IKER, UMR5478, 64100 Bayonne
(2) LIASD - Université Paris 8, 2 rue de l'Université, 93526 Saint-Denis, France
(3) LIFO - Université d'Orléans, 6 Rue Léonard de Vinci, 45067 Orléans, France
(4) Université Paris Nanterre, MoDyCo, 92000 Nanterre
melanie.jouitteau@iker.cnrs.fr, am@up8.edu,
jean-yves.antoine@univ-tours.fr, loic.grobol@gmail.com

KEYWORDS : data acquisition, participatory sciences, minorized languages, Breton, ASR

1 Introduction

The accelerated technologization of human relationships (Sayers *et al.*, 2021) endangers the practice of languages for which new linguistic technologies cannot be deployed. Building digital resources that can be used in NLP is therefore an essential task in preserving human linguistic diversity. As far as oral technologies are concerned, there are software solutions for the participative acquisition of data, such as Common Voice (Ardila *et al.*, 2020). It is however clear that appropriation remains insufficient by speaking communities [1]. We present here a pilot project in the Breton linguistic context. It aims at validating the following hypothesis : early interdisciplinary collaboration with speaking communities for the design of data acquisition tools significantly increases their appropriation and therefore the effectiveness of the tools. Concretely, the project YAR [*Yezh Ar vRo* - the language of the country] proposes to develop two tools :

1. A mobile application for geolocalized speech collection (YAR-app)
2. A web platform for participatory transcription (`yar.bzh`)

Both are meant to answer the specific needs for tools expressed by civil society. In the first stage, we gather pre-existing oral corpora and enrich them with metadata including geolocation. We develop the mobile application YAR-app for collecting geolocated sound and test it. In the second stage, we transcribe existing archives to pre-populate the map

*. This paper was originally published in French in the Proceeding of the LIFT2 conference in Orléans, 2024, Nov 14-15.

1. Common Voice has made it possible to collect 25 hours of aligned oral Breton in 5 years (2019-2024).

with proximity data. We develop the web platform `yar.bzh` for participatory transcription. The transcribed corpora is used as a base to develop of pedagogical functions in yar.bzh. Educational tools whose purpose is transcription both use and provide aligned corpus that in turn constitutes the NLP resource. Finally, we develop a pre-transcription solution assisted by automatic speech recognition (ASR). ASR will thus support the growth of the volume of transcribed material to be transformed into pedagogical transcription exercises and corrected by humans, incrementally building a gold standard corpus.

2 An integrated participatory approach

2.1 Involvement of civil society

The appropriation by speakers of data acquisition tools as explored in Millour (2020) can be achieved through early and ongoing involvement of stakeholders in the speaking community : speakers, language workers (translators, teachers, learners, linguists, collectors, archivists) and representatives of local community cultural and economic interests. Beyond academics, YAR mobilizes in Brittany the Dastum network of collectors of the Breton language, an educational hub of three associations teaching Breton to adults (Roudour, Stumdi, Mervent), and a translation company for the dubbing industry. Finally, an endowment fund for the development of technologies for the automatic processing of Breton, Bretagne Numérique, provides a link between the industrial and associative worlds and supports the approach by organizing datathons serving the project. These key players in civil society are co-designers, promoters and end users of digital tools. The integrated collaborative approach defines which data acquisition tools will serve this community, independently of the need to build resources for NLP.

2.2 Another bridge between NLP and Human sciences is possible

The integrated participatory approach requires technological expertise (including ergonomics), dialectological knowledge and community engagement. This outlines a space of collaboration space between NLP and Human sciences that clearly extends beyond the "digital humanities" and NLP for linguistics. Colleagues from educational sciences, fieldwork descriptive linguistics, sociology of language, literature, anthropology or ethnology can be embody an effective point of contact for NLP with language policy structures, public and private language teaching structures, local archives (written and audio), and the local economic stakeholders (regional newspapers, museums, tourist offices, industries with a local image, etc.).

3 Challenges

3.1 Social challenges

The Breton ecosystem includes less than 200 000 speakers (Broudic, 2009), all bilinguals with French, mostly acculturated to digital tools available in French. Public policies describe a need for digital tools — and in particular ASR systems — on the one hand, and educational resources including sound on the other (Tyers & Howell, 2021; Ropers, 2007). YAR meets these needs by equipping data collection for NLP. It also provides resources for the creation of teaching materials, and helps learners socialize in Breton and acquire dialectal flexibility.

The involvement of transcription experts, the use of pre-existing corpora for priming and the distribution of datathons across the linguistic area ensure the quality, quantity and diversity of the data. The transcription is based on standard spelling and its dialectal variants, all relatively well established within the community.

The integration of geolocation makes it possible to anchor learning paths in the territory by mapping dialectal diversity. It also makes the language visible in the public space, and offers digital support for local authority projects and extended uses of augmented tourism.

3.2 Technological issues

The adequacy of the solutions developed to the real needs of the linguistic community ensures that the data collected will be representative of the oral uses of the language in its dialectal diversity and of speaker profiles. These data join the inventory of available aligned data and support the development of an ASR system. Our exploratory, iterative and state-of-the-art approach to automatic speech recognition in Breton (Duval-Guennoc, 2022 présent), ensures the development of a high-performance system.

3.3 Ergonomic issues

The appropriation of technologies is a delicate issue in which psychological, sociological or economic factors intervene. The field of disability assistance is for instance full of examples of technical aids providing an objective benefit to people but abandoned for psychological or social representation reasons. In all cases, the appropriation of a digital application by its used cannot been expected if it is not as user-friendly as possible. In YAR, this requirement is achieved by implementing a user-centered design approach involving representative users throughout the application life cycle. In particular, the needs analysis and ideation stages will involve collaborative brainstorming sessions in the form of focus groups with (1) local associations promoting socialization in the language, and (2 - for the transcription application) with several educational centers that teach the language to adults.

4 Conclusion

We study the impact of involving speaking communities in the construction of technological tools for the preservation and documentation of their languages through an integrated participatory approach. The appropriation of data acquisition tools depends directly on our ability to serve these communities according to their point of view and independently of the needs of NLP. This fundamenNLP research on data acquisition methods in minority language ecosystems thus proposes to instrumentalize the construction of tools to serve communities.

We are looking for collaborations in order to experiment with this participatory model on other socio-linguistic ecosystems. In the French state alone, more than fifty languages have digitally equipped speakers (access to electricity, internet, smartphones), scarce NLP resources and a documented social demand, like in Kanaky or French Guiana. The geolocation function that we propose for YAR-app could be of particular interest for the visibility of so-called non-territorial languages such as Western Armenian, Erromintxela, Kaló, Rromani or Sintó.

Références

ARDILA R., BRANSON M., DAVIS K., KOHLER M., MEYER J., HENRETTY M., MORAIS R., SAUNDERS L., TYERS F. & WEBER G. (2020). Common voice : A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4218–4222, Marseille, France : European Language Resources Association. DOI : 10.48550/arXiv.1912.06670.

BROUDIC F. (2009). *Parler breton au XXIe siècle : Le nouveau sondage de TMO Régions.* Emgleo Breiz.

DUVAL-GUENNOC G. (2022-présent). Anaouder, a vosk model for the breton lanugage. https://github.com/gweltou/our-voices-model-competition/tree/vosk-br/.

MILLOUR A. (2020). *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées*. Thèse de doctorat, Sorbonne Université.

ROPERS C. (2007). KYG : A corpus of spoken breton for both researchers and advanced learners. *Journal of Celtic Language Learning*, **5-24**.

SAYERS D., SOUSA-SILVA R., HÖHN S., AHMEDI L., ALLKIVI-METSOJA K., ANASTASIOU D., BEŇUŠ Š., BOWKER L., BYTYÇI E., CATALA A., ÇEPANI A., CHACÓN-BELTRÁN R., DADI S., DALIPI F., DESPOTOVIC V., DOCZEKALSKA A., DRUDE S., FORT K., FUCHS R., GALINSKI C., GOBBO F., GUNGOR T., GUO S., HÖCKNER K., LÁNCOS P., LIBAL T., JANTUNEN T., JONES D., KLIMOVA B., KORKMAZ E., MAUČEC MIRJAM S., MELO M., MEUNIER F., MIGGE B., MITITELU VERGINICA B., NÉVÉOL A., ROSSI A., PAREJA-LORA A., SANCHEZ-STOCKHAMMER C., ŞAHIN A., SOLTAN A., SORIA C., SHAIKH S., TURCHI M. & YILDIRIM YAYILGAN S. (2021). The

Dawn of the Human-Machine Era : A forecast of new and emerging language technologies. Report brings together insights from specialists in the fields of language technology and linguistic research., HAL : hal-03230287.

TYERS F. M. & HOWELL N. (2021). Morphological analysis and disambiguation for breton. *Language Resources and Evaluation*, **55**(2), 431–473. DOI : 10.1007/s10579-020-09510-8.