Construire l'appropriation des applications de collecte

Mélanie Jouitteau[1] Loïc Grobol[2,3] Alice Millour[4] Jean-Yves Antoine[5,6]
(1) IKER, CNRS, UMR 5478, Université de Pau et des Pays de l'Adour et
Université Bordeaux Montaigne, Bayonne, 64100, France
(2) MoDyCo, Université Paris Nanterre, UMR 7114, 92001, Nanterre, France
(3) Lattice – École Normale Supérieure, 92120, Montrouge, France
(4) LIASD (Équipe Pastis) - Université Paris 8 Vincennes Saint-Denis, 93526, Saint-Denis, France
(5) LIFAT - Université de Tours, 37200, Tours, France
(6) LIFO - Université d'Orléans, 45067, Orléans, France
melanie.jouitteau@iker.cnrs.fr, loic.grobol@gmail.com,
am@up8.edu, jean-yves.antoine@univ-tours.fr

MOTS-CLÉS : acquisition des données, sciences participatives, langues minorisées, breton, ASR.

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

1 Introduction

En contexte de technologisation accélérée des rapports humains (Sayers et al. 2021), les langues pour lesquelles les nouveaux outils ne peuvent être déployés risquent des baisses de pratique potentiellement fatales. Construire des ressources numériques utilisables en TAL devient une tâche essentielle de préservation de la diversité linguistique humaine. Il existe des solutions logicielles pour l'acquisition de données (Ardila *et al.*, 2020), mais elles rencontrent une appropriation insuffisante par les communautés parlantes [1]. Nous présentons un projet pilote d'envergure déjà significative visant à valider l'hypothèse de recherche qu'une collaboration interdisciplinaire précoce avec les communautés parlantes pour la conception des outils d'acquisition des données augmente significativement leur appropriation et donc leur efficacité. Concrètement, le projet YAR [*Yezh Ar vRo* - la langue du pays] propose de développer deux outils d'acquisition de données :

1. Une application mobile de collecte de parole géolocalisée (YAR-app)
2. Une plateforme web de transcription participative (yar.bzh)

Dans un premier temps, nous rassemblons des corpus oraux préexistants, et les enrichissons en métadonnées, dont la géolocalisation. Nous développons l'application mobile YAR-app de collecte de son géolocalisé et la testons. Dans un second temps, nous outillons la transcription des corpus oraux. Nous transcrivons un fond d'amorçage pour pré-peupler la carte en données

1. Common Voice a permis de rassembler 25h de breton oral aligné en 5 ans (2019-2024).

de proximité. Nous développons la plateforme web yar.bzh pour la transcription participative. Les outils pédagogiques dont l'objet est la transcription sont alimentés par la collecte de données et fournissent du corpus aligné qui constitue la ressource TAL. Nous explorons une solution de pré-transcription assistée par reconnaissance vocale automatique (ASR) qui facilite incrémentalement la transcription.

2 Une démarche participative intégrée

L'appropriation par la foule des outils d'acquisition de données telle qu'explorée dans Millour (2020) peut être obtenue par une implication précoce et continue des tenants de la communauté parlante : locuteurs, travailleurs de la langue (traducteurs, pédagogues, apprenant.es, linguistes, collecteurs, archivistes) et représentants des intérêts communautaires culturels et économiques locaux. Au-delà des universitaires, YAR mobilise en Bretagne le réseau Dastum de collecteurs de la langue bretonne, un pôle pédagogique de trois associations d'enseignement du breton pour adultes (Roudour, Stumdi, Mervent), et une entreprise de traduction pour l'industrie du dtoublage. Enfin, un fond de dotation pour le développement des technologies du traitement automatique du breton, Bretagne Numérique, fait le lien entre le monde industriel et associatif et soutient la démarche par l'organisation de datathons servant le projet. Ces acteurs clef de la société civile sont co-concepteurs, promoteurs et utilisateurs finaux des outils numériques. La démarche collaborative intégrée définit quels outils d'acquisition de données vont servir cette communauté indépendamment des besoins de constitution de ressources pour le TAL.

2.1 Un autre pont TAL/SHS est possible

La démarche participative intégrée nécessite expertise technologique (ergonomie incluse), connaissance dialectologique et engagement communautaire. L'espace de collaborations TAL-SHS que cela dessine déborde nettement les « humanités numériques » et la linguistique outillée. Les collègues de sciences de l'éducation, linguistiques de terrain, sociologie du langage, littérature, anthropologie ou ethnologie peuvent être pour le TAL un point de contact et levier efficace vers les structures de politique linguistiques, les structures publiques ou privées d'enseignement des langues, les archives locales écrites ou sonores, le tissu économique local (journaux régionaux, musées, offices du tourisme, industries à image de terroir, etc.).

3 Paramètres du domaine empirique

L'écosystème breton comprend moins de 200.000 locuteurs (Broudic, 2009), tous bilingues français-breton, acculturés aux outils numériques en français. Les politiques publiques identifient clairement un manque d'outils numériques adaptés en matière d'ASR et de ressources pédagogiques incluant du son, lié au manque de ressources numériques associées (Tyers & Howell, 2021; Ropers, 2007). L'intégration de la géolocalisation dans YAR-app permet de cartographier la diversité dialectale et de créer des parcours d'apprentissage ancrés dans le territoire. Elle visibilise la langue dans l'espace public, offre un support numérique aux projets des collectivités locales à l'échelle communale, et aux usages étendus de tourisme augmenté. Elle facilite la socialisation des apprenants dans la langue bretonne en outillant la collecte de proximité, fournit des supports pédagogiques prêts-à-créer, adaptés à leurs porteurs pédagogiques et au besoin d'apprentissage de flexibilité dialectale à l'écoute. La transcription peut s'appuyer sur un standard établi et sa flexibilité relativement installée pour la graphie de formes dialectales. L'implication d'experts en transcription, le recours à des corpus préexistants pour l'amorçage et trois points de datathons répartis sur l'aire parlante assurent la qualité, la quantité et la diversité des données collectées et transcrites.

4 Enjeux technologiques en TAL

L'adéquation des solutions développées aux besoins réels de la communauté linguistique assure que les données collectées seront représentatives des usages oraux de la langue dans sa diversité dialectale et de profils de locuteurs. Ces données rejoignent l'inventaire des données alignées disponibles et soutiennent le développement d'un système d'ASR. Notre approche exploratoire et itérative, informée de l'état de l'art sur cette langue (Duval-Guennoc, 2022 présent), assure le développement d'un système performant pour une langue peu dotée.

5 Enjeux ergonomiques : conception centrée utilisateur

L'appropriation des technologies est une question délicate dans laquelle interviennent des facteurs d'ordre psychologiques, sociologiques ou économiques. On a observé par le passé des détournements d'usages de la technologie (exemple des SMS), tandis que le domaine du handicap regorge d'exemples d'aides techniques apportant un bénéfice objectif aux personnes mais délaissées pour des raisons psychologiques ou d'image sociale. Dans le cas du numérique, une condition nécessaire est toutefois incontournable : on ne peut s'attendre à l'appropriation d'une application si celle-ci n'est pas la plus conviviale possible. Dans YAR, cette exigence est atteinte par la mise en place d'une démarche de conception centrée-utilisateur associant des personnes utilisatrices représentatives tout au long du cycle de

vie logiciel des application. En particulier, les étapes d'analyse des besoins et d'idéation mettront en jeu des séances de brainstorming collaboratif sous forme de focus groups avec des associations de sociabilisation dans la langue, puis pour l'application de transcription avec un pôle pédagogique d'enseignants pour adultes.

6 Conclusion

Nous étudions l'impact de l'implication des communautés parlantes dans la construction d'outils technologiques pour la préservation et la documentation de leurs langues, avec l'hypothèse que cette implication peut être construite par une démarche participative intégrée. Cette recherche fondamentale sur les modes d'acquisition de données en écosystème de langue minorisée propose d'instrumentaliser la construction pragmatique d'outils au service des communautés. L'appropriation des outils d'acquisition de données y dépend directement de notre capacité à servir ces communautés selon leur point de vue et indépendamment des besoins du TAL. Nous cherchons des collaborations ouvrant au test du modèle participatif sur d'autres écosystèmes socioliguistiques. Plus d'une cinquantaine de langues de France ont des locuteurs numériquement équipés (électricité, internet, smartphones), de maigres ressources TAL et une demande sociale documentée, comme les langues kanak ou de Guyane française. La fonction de géolocalisation que nous proposons pourrait intéresser particulièrement la visibilisation des langues dites non-territoriales comme l'arménien occidental, l'erromintxela, le kaló, le rromani ou le 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**.

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.