Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

LiLA : Outil d'augmentation automatisée des données vocales participatives de Lingua Libre

Mathilde Hutin[1] Marc Allassonnière-Tang[2] Lucas Prégaldiny[3] Lucas Lévêque[3]
(1) Université de Lorraine, CNRS, ATILF, 44 av. de la Libération, F-54000 Nancy, France
(2) Université de Paris Cité, MNHN, CNRS, Laboratoire EA (UMR 7206), 17 place du Trocadéro, 75016 Paris, France
(3) Contributeur Wikimédia, France
mathilde.hutin@cnrs.fr, marc.allassoniere-tang@mnhn.fr, lucas@lugnumerique.fr

RESUME

La constitution de corpus vocaux, nécessaires à l'exploration de la phonétique et de la phonologie des langues du monde, soulève de nombreux défis. La constitution de corpus multi-dialectes, permettant d'explorer la variation dialectale, ou de corpus multilingues, permettant de comparer plusieurs langues, est d'autant plus difficile que, pour que chaque dialecte/langue soit comparable aux autres dans le corpus, les données doivent avoir été enregistrées dans les mêmes conditions (même matériel, même protocole…). Une solution à ces défis semble envisageable aujourd'hui grâce aux données participatives, par définition administrées et enregistrées par des volontaires, et donc moins coûteuses à tous points de vue pour la communauté scientifique. En mars 2025, Lingua Libre, la médiathèque linguistique participative de Wikimédia France ouverte depuis 2018, compte ~1,4M enregistrements en 284 langues par 2.547 individus à travers le monde : notre projet est de créer un outil pour rendre ces données brutes exploitables par les linguistes.

ABSTRACT

LiLA: A Tool for the Automatic Augmentation of Crowd-sourced Vocal Data from Lingua Libre.
Creating voice corpora, although necessary to explore the phonetics and phonology of the world's languages, is a challenging task. Multi-dialect or multilingual corpora, allowing to investigate dialectal or typological variation, are all the more difficult to create that, for each dialect/language to be comparable to the others in the corpus, the data must have been recorded in similar conditions (same material, same protocol…) across the world. A solution to this challenge can be envisioned today thanks to crowd-sourced data, by definition administered and recorded by volunteers and therefore less expensive in both time and money for researchers. In March 2025, Lingua Libre, Wikimedia France's participative linguistic library opened in 2018, gathers ~1.4M recordings in 284 languages by 2,547 speakers across the world: Our project aims to create a tool to turn its raw audio data into data exploitable by linguists.

MOTS-CLES : Lingua Libre, Wikimédia, Données participatives, Phonétique, Phonologie, Typologie.
KEYWORDS : Lingua Libre, Wikimedia, Crowd-sourced data, Phonetics, Phonology, Typology

ARTICLE : **Accepté à** Atelier ParCoL : Science Participative pour les Données et Corpus Linguistiques.

La constitution de corpus vocaux, nécessaires à l'exploration de la phonétique et de la phonologie des langues du monde, soulève un certain nombre de défis, par exemple logistiques (recrutement des participants, organisation des récoltes, déplacements…), et implique de nombreux coûts (financiers, humains, carbone…). La constitution de corpus multi-dialectes, permettant d'explorer la variation dialectale, et plus encore de corpus multilingues, permettant de comparer plusieurs langues, voire plusieurs familles de langues, est d'autant plus difficile que, pour que chaque dialecte ou langue soit comparable aux autres dans le corpus, les données doivent avoir été enregistrées avec le même matériel, dans les mêmes conditions, suivant le même protocole, et par des cohortes similaires. Une solution à ces défis semble envisageable aujourd'hui grâce aux données participatives.

Les données participatives sont des données, en l'occurrence vocales, issues d'initiatives citoyennes, qui sont, par définition, administrées et enregistrées par des volontaires, et donc moins coûteuses à tous points de vue pour la communauté scientifique. Ici nous utilisons les données de Lingua Libre (LiLi, https://lingualibre.org/), la médiathèque linguistique participative de Wikimédia France, ouverte depuis 2018. En mars 2025, LiLi compte ~1,4M enregistrements en 284 langues par 2 547 individus à travers le monde (cf. Fig. 1). Des études ont montré que les données de LiLi sont en partie comparables aux données enregistrées par des professionnels (Hutin & Allassonnière-Tang, 2022a), et peuvent être utilisées pour explorer des questions de typologie (Hutin & Allassonnière-Tang, 2022bcd) ou de variation dialectale (Hutin & Allassonnière-Tang, 2023), et ce par tout un chacun, puisque les données sont hébergées sous licence libre sur Wikimedia Commons.

FIGURE 1: Répartition géographique des contributeurs-contributrices à Lingua Libre en 2024.

Notre projet vise à développer LiLA (Lingua Libre Augmenté), un outil inspiré de travaux passés (Hutin & Allassonnière-Tang, 2022bcd, 2023) pour augmenter les données vocales de LiLi avec (i) des métadonnées sur les locuteurs-locutrices en provenance de LiLi, (ii) des informations sur la langue (famille, nombre de locuteurs-locutrices) en provenance de Wikidata, et (iii) des alignements au niveau du mot et au niveau du phone grâce à WebMAUS (Schiel, 1999 ; Kisler et al., 2017). Comme l'alignement est spécifique à chaque langue, le projet est limité pour l'instant aux langues à la fois présentes dans LiLi et disponibles sur WebMAUS, soit 26 langues (cf. Tab 1).

Langues WebMAUS	Enregistrements LiLi	Locuteurs-locutrices LiLi
Français	406 312	657
Polonais	95 209	43
Anglais	79 704	228
Roumain	22 989	7
Catalan	22 888	18
Basque	20 258	131
Allemand	19 108	72
Espagnol	16 579	84
Russe	14 029	45
Italien	11 075	17
Suédois	9 179	15
Arabe (macro)	7 321	22
Norvégien	5 412	7
Persan	4 117	7
Afrikaans	2 944	4
Néerlandais	1 764	11
Luxembourgeois	916	2
Japonais	895	9
Géorgien	880	3
Alémanique	744	3
Maltais	684	2
Thai	289	4
Hongrois	236	4
Finnois	219	4
Albanais	88	9
Islandais	19	3

TABLE 1 : Listes des langues disponibles sur WebMAUS qui ont un nombre d'enregistrements non-nul sur LiLi, avec le nombre d'enregistrements et le nombre de locuteurs-locutrices dans LiLi en date du 25 mars 2025. Les langues sont classées par nombre décroissant d'enregistrements.

LiLA est une interface web développée en Python qui permet de générer un ensemble de fichiers : (i) les fichiers-sons, (ii) les métadonnées des locuteurs-locutrices, (iii) les alignements au niveau du mot et au niveau du phone (ex. Fig. 2), et (iv) un tableur (ex. Tab. 2).

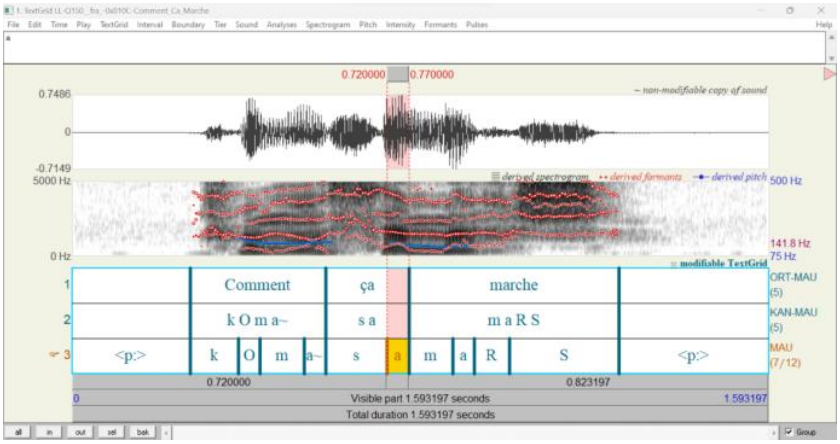

FIGURE 2 : Exemple d'alignement sous forme de textgrid dans Praat (Boersma & Weenink, 2024) généré par WebMAUS pour le fichier-son de LiLi *LL-Q150_fra.-0x010C_Comment_Ca_Marche*. L'audio « Comment ça marche » est aligné au niveau du mot en orthographe du français sur le premier tier, en transcription phonétique sur le deuxième, et au niveau du phone sur le troisième.

phone	phone_ duration	word_ ortho	word _api	preceding _phone	following _phone	speaker _ID	speaker _gender	speaker_ location	speaker _L1	language	language _size	language _family
s	0.217	ça	sa	#	a	Mathsou	femme	Paris	français	français	113 000 000	IE
a	0.149	ça	sa	s	#	Mathsou	femme	Paris	français	français	113 000 000	IE

TABLE 2 : Exemple fictif du tableur généré par LiLA. Chaque ligne correspond à un phone du signal. Les colonnes indiquent le phone (en SAMPA ; Wells, 1997), sa durée (en seconde), son orthographe dans la langue, sa transcription en SAMPA, le phone précédent, le phone suivant, l'identifiant du locuteur-locutrice, son genre (tel qu'auto-déclaré), son lieu de résidence, sa langue maternelle, la langue du mot prononcé, son nombre de locuteurs-locutrices et sa famille phylogénétique.

En somme, nous souhaitons proposer un outil pour transformer des données brutes disponibles en ligne en corpus structuré exploitable par les linguistes. Cette preuve de concept permet d'envisager, à terme, la publication d'un grand corpus multilingue sous licence libre créé d'après cette méthodologie mais de surcroît corrigé manuellement.

Références

BOERSMA P. & WEENINK D. (2024). Praat: doing phonetics by computer [Computer program]. Version 6.4.25, retrieved 8 December 2024 from http://www.praat.org/

HUTIN M. & ALLASSONNIERE-TANG M. (2022a). Crowd-sourcing for Less-resourced Languages: Lingua Libre for Polish. *Proceedings of SIGUL 2022*. Marseille, France, 24-25 juin 2022, 41-47. HAL : ⟨hal-03706257⟩

HUTIN M. & ALLASSONNIERE-TANG M. (2022b). Investigating phonological theories with crowd-sourced data: The Inventory Size Hypothesis in the light of Lingua Libre. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics (ACL). Seattle, USA, July 14th, 2022, pp. 23-28. DOI: https://doi.org/10.18653/v1/2022.sigmorphon-1.3

HUTIN M. & ALLASSONNIERE-TANG M. (2022c). Languages Worldwide and the World Wide Web: Crowdsourcing on the Internet to Explore Linguistic Theories. *Diversity of Methods and Materials in Digital Human Sciences: Proc. of the Digital Research Data and Human Sciences DRDHum Conference 2022*, Dec. 2022, Jyväskylä, Finland: U of Jyväskylä. 136-147. HAL: ⟨hal-03887378⟩

HUTIN M. & ALLASSONNIERE-TANG M. (2022d). Operation LiLi: Using crowd-sourced data and automatic alignment to investigate the phonetics and phonology of less-resourced languages. *Languages* 7: 234, *Advances in Phonetic Sciences: Role of Speech Corpora and Automatic Processing* (special issue). DOI: https://doi.org/10.3390/languages7030234

HUTIN M. & ALLASSONNIERE-TANG M. (2023). L'apport des données participatives pour l'étude linguistique des français du monde: le cas de l'opposition /a~ɑ/. *Journal of French Language Studies*, 1-24. DOI: https://doi.org/10.1017/S0959269523000200

KISLER T., REICHEL U. & SCHIEL F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language* 45: 326–47. DOI: https://doi.org/10.1016/j.csl.2017.01.005

SCHIEL F. 1999. Automatic phonetic transcription of non-prompted speech. *14th International Congress of Phonetic Sciences: ICPhS 99*, San Francisco, CA, USA, Aug. 1-7; 607-10.

WELLS J. (1997). *SAMPA Computer Readable Phonetic Alphabet*. Vol. Part IV. Berlin: Mouton de Gruyter.