CoRéFO 0.1: Premières explorations pour le Corpus de Référence du Français Oral

Mathilde Hutin[1] Marc Allassonnière-Tang[2]
(1) Université de Lorraine, CNRS, ATILF, 44 avenue de la Libération, F-54063, Nancy, France
(2) U Paris-Cité, MNHN, CNRS, Eco-anthropologie, 17 place du Trocadéro, 75016 Paris, France
mathilde.hutin@cnrs.fr, marc.allassonniere-tang@mnhn.fr

MOTS-CLÉS : Corpus oral, français, données participatives, Common Voice.

KEYWORDS: Oral corpus, French, crowd-sourced data, Common Voice.

1 Introduction

Les technologies de traitement de la parole présentent aujourd'hui encore trois défauts majeurs. D'abord, elles favorisent massivement un petit nombre de langues parlées par un grand nombre de personnes (les langues dites "bien dotées"), au détriment de langues parlées par des communautés plus petites (les langues dites "peu dotées"). Ensuite, même pour les premières, les représentations linguistiques sont souvent monolithiques, c'est-à-dire que les technologies traitent bien la parole dite standard, mais peinent à prendre en compte la variation, par exemple dialectale. Enfin, même dans le cas de parole standard dans une langue largement parlée, certaines technologies de pointe, tant dans le domaine public que dans l'industrie privée, sont encore relativement insatisfaisantes.

Le projet CoCLiCo (Common Collection of Linguistic Corpora) vise à constituer une méthodologie pour doter de nombreuses langues de corpus vocaux de référence, suffisamment propres et fiables pour servir à la fois de bases de données à la recherche en linguistique théorique, et de données d'entraînement finement annotées pour les technologies du langage. Pour des raisons pratiques, la première phase de ce développement se concentre sur le français. Ce sous-corpus (CoRéFO) sera composé de fichiers-sons (en format .mp3) avec leur transcription orthographique (en format .txt), leur alignement au niveau du mot et au niveau du phone (en format .TextGrid) et de leur métadonnées (en format .xlsx).

2 Données et méthode

Les fichiers audios et les transcriptions proviennent de Common Voice (Ardila *et al.*, 2020), la plateforme libre de Mozilla pour la création de données par la communauté. La plateforme

propose un dispositif d'enregistrement de lecture de phrases. Les données sont donc représentatives essentiellement de parole continue lue. Le projet s'étendra plus tard à d'autres styles de parole, puisque Common Voice vient de s'élargir aux données de parole spontanée.

Les enregistrements ont été filtrés pour que ne soient téléchargées que les données (i) validées par la communauté Common Voice et (ii) pour lesquelles le genre, l'âge et le dialecte du locuteur-locutrice sont renseignés. Le total s'élève à 214 350 paires fichier-son (au format .mp3) + transcription (au format .txt). Cela correspond à 214 350 phrases, soit 5 417 527 mots, dont 400 129 mots uniques.

Les fichiers-son et leur transcription ont été téléchargés depuis Common Voice le 20 juillet 2025. Courant juillet 2025, à peu près la moitié des fichiers a été alignée avec WebMAUS (Schiel, 1999, 2004), la version web du Munich Automatic Segmentation System MAUS. Cet aligneur a été choisi car des études ont montré que WebMAUS est aussi performant qu'un alignement manuel pour de la parole continue en allemand (Wesenick & Kipp, 1996; Kipp *et al.*, 1996) et les frontières posées par WebMAUS-français sur des données de lecture de mots isolés s'écartent de la correction manuelle par une experte locutrice native du français de ∼0,01 seconde (Hutin & Allassonnière-Tang, 2023).

Les alignements générés sont des textgrids lisibles par Praat (Boersma & Weenink, 2024). Les textgrids ont le format suivant :

— tier 1 = alignement au niveau du mot en orthographe du français,
— tier 2 = alignement au niveau du mot en transcription SAMPA (Wells, 1997),
— tier 3 = alignement au niveau du phone en SAMPA.

Enfin, 1 000 alignements ont été corrigés manuellement par un étudiant en linguistique. Cet étudiant n'étant pas locuteur natif du français, 132 alignements ont aussi été corrigés par la première autrice, locutrice native du français et phonéticienne entraînée afin de pouvoir calculer un score d'accord inter-annotateur. Ces corrections ont mis en lumière plusieurs problèmes (détaillés en 3) auxquels il faudra remédier pour la prochaine étape de constitution du CoRéFO. Les données de CoRéFO 0.1 sont disponibles à https://osf.io/8f7x4/.

3 Résultats

Cette phase de test a permis de produire 1 132 alignements manuels (dont 264 en double) pour 1 000 enregistrements. Les 1 000 enregistrements alignés correspondent à 8 754 mots prononcés par 26 locuteurs, tous des hommes. Parmi ces enregistrements, 2,1% ont été produits par des personnes étiquetées "teens", 5,8% par des personnes dans la vingtaine, 50,8% dans la trentaine, 32,0% dans la quarantaine, 5,6% dans la cinquantaine et 3,6% dans la soixantaine. Pour ce qui est de la variété de français, 98,3% sont des enregistrements de français de France, 0,7% de Suisse, 0,5% de Belgique, et 0,4% d'Allemagne.

Surtout, cette phase de test a mis en évidence un certain nombre de points d'attention qui devront être anticipés lors de la création du corpus final.

D'abord, concernant le tri des données de Common Voice, il est apparu que certains fichiers sont encore de mauvaise qualité, par exemple avec beaucoup de bruit de fond ou des grésillements dus à un mauvais micro. Une solution serait de ne conserver que les enregistrements qui n'ont jamais récolté de "downvotes" lors de la validation par les autres internautes. L'autre solution serait que, puisque nous envisageons de corriger à la main tous les enregistrements, les audios de mauvaise qualité soient identifiés et supprimés de la base de données lors de la correction manuelle. Les deux solutions ne sont d'ailleurs pas mutuellement exclusives.

Toujours concernant le filtrage des données, il est apparu dans des travaux récents (Zhang *et al.*, 2025) que les "client_id", c'est-à-dire les identifiants-clients rattachés aux métadonnées des locuteurs-locutrices, renvoient parfois à un compte partagé par plusieurs personnes, avec donc des caractéristiques socio-démographiques différentes. La solution envisagée pour assurer la correspondance entre voix et métadonnées consiste à croiser la liste de nos fichiers filtrés depuis Common Voice avec la liste de fichiers identifiés comme problématiques par Zhang *et al.* (2025), disponible sur la page Hugging Face du projet Vox Communis.

La piètre qualité des alignements avec WebMAUS s'est aussi avérée problématique. Après enquête, il est apparu qu'une des causes était l'encodage de la transcription. Ce problème sera résolu en corrigeant automatiquement les caractères qui ont posé le plus de problèmes. Cependant, nous envisageons également de conduire une étude comparative entre WebMAUS et SPPAS (Bigi, 2015) préalablement à l'établissement du corpus final. Le but sera notamment d'identifier lequel des deux outils permet de réduire le plus le temps de correction manuelle.

4 Conclusion et discussion

Nous avons produit filtrages, alignements automatiques et corrections manuelles pour 1000 enregistrements francophones provenant de Common Voice. Cette étude-pilote a servi à baliser le terrain pour constituer un corpus de référence du français oral, constitué de fichiers-son dont la qualité sera attestée, de transcriptions proprement encodées et de textgrids intégralement corrigés manuellement. Les problèmes identifiés ont permis de prévoir des garde-fous pour assurer la faisabilité et la qualité du corpus final, dont il est prévu qu'il comprenne, au moins dans un premier temps, ∼4500 triplets enregistrement-transcription-alignement, soit à peu près 4 heures de parole dans une quinzaine de variétés de français.

Notre espoir est de développer une méthodologie de collecte et de traitement pour que la communauté scientifique puisse créer à moindre coût une base de données massive qui permette l'exploration précise et fiable de la phonétique et de la phonologie d'un nombre toujours grandissant de langues, dans des styles de parole diversifiés, par des cohortes aussi équilibrées que possible.

Remerciements

Nous remercions chaleureusement Punsisi Liyanage, étudiant du master Erasmus Mundus EMLex, qui a nettoyé manuellement 1000 textgrids lors de son stage à l'ATILF en 2025.

Références

ARDILA R., BRANSON M., DAVIS K., HENRETTY M., KOHLER M., MEYER J., MORAIS R., SAUNDERS L., TYERS F. M. & WEBER G. (2020). Common Voice : A Massively-Multilingual Speech Corpus. In *Proceedings of LREC*.

BIGI B. (2015). SPPAS - MULTI-LINGUAL APPROACHES TO THE AUTOMATIC ANNOTATION OF SPEECH. *The Phonetician. Journal of the International Society of Phonetic Sciences*, **111-112**(ISSN :0741-6164), 54–69. HAL : hal-01417876.

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

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

HUTIN M. & ALLASSONNIÈRE-TANG M. (2023). L'apport des données participatives pour l'étude linguistique des français du monde : le cas de l'opposition /a/. *Journal of French Language Studies*, **34**(2), 249–272. DOI : 10.1017/S0959269523000200.

KIPP A., WESENICK M.-B. & SCHIEL F. (1996). Automatic detection and segmentation of pronunciation variants in German speech corpora. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 1, p. 106–109 vol.1. DOI : 10.1109/ICSLP.1996.607048.

SCHIEL F. (1999). Automatic phonetic transcription of non-prompted speech. In J. J. OHALA, Éd., *Proceedings of the XIVth International Congress of Phonetic Sciences : ICPhS 99 ; San Francisco, 1 - 7 August 1999*, p. 607 –610, San Francisco.

SCHIEL F. (2004). MAUS goes iterative. In M. T. LINO, M. F. XAVIER, F. FERREIRA, R. COSTA & R. SILVA, Éds., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA).

WELLS J. (1997). *SAMPA computer readable phonetic alphabet*, In D. GIBBON, R. MOORE & R. WINSKI, Éds., *Handbook of Standards and Resources for Spoken Language Systems*, volume Part IV. Mouton de Gruyter.

WESENICK M.-B. & KIPP A. (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 1996)*, p. 129–132. DOI : 10.21437/ICSLP.1996-33.

ZHANG M., FARHADIPOUR A., BAKER A., MA J., PRICOP B. & CHODROFF E. (2025). Quantifying and Reducing Speaker Heterogeneity within the Common Voice Corpus for Phonetic Analysis. In *Interspeech 2025*, p. 3933–3937. DOI : 10.21437/Interspeech.2025-2027.