Quelles données pour l'exploration de la variation dialectale en français?

Etude de cas sur les voyelles ouvertes du français

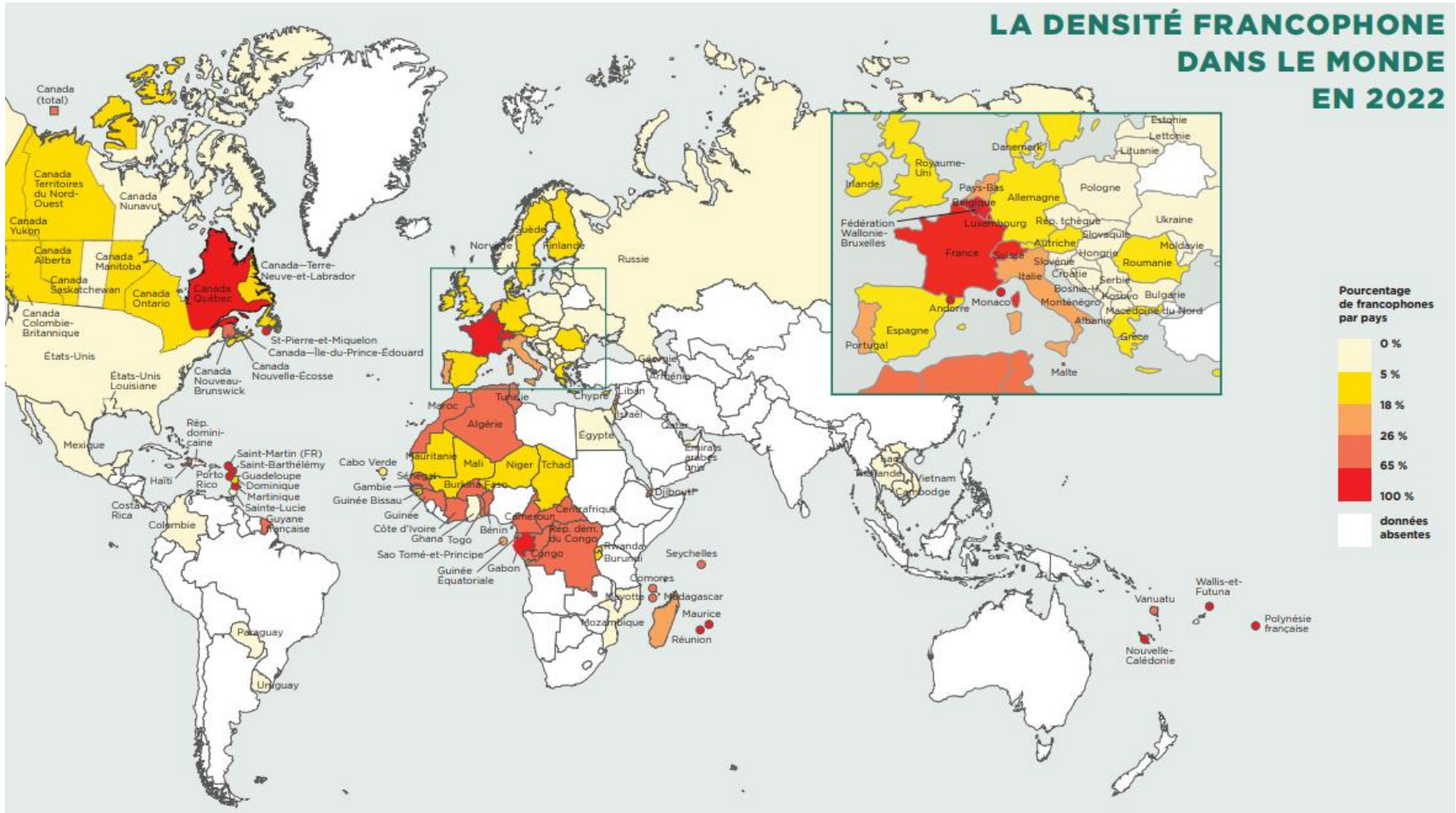
Mathilde Hutin

F.R.S.-FNRS – Université catholique de Louvain, Institut Langage et Communication

Journées (Inter-)Phonologie du Français Contemporain

Paris – 29 novembre 2024

Le français dans le monde



La densité francophone dans le monde en 2022. In *La langue française dans le monde. Synthèse 2022*. Organisation internationale de la Francophonie. P. 9.

- Langue romane occidentale
- ~320M de locuteurs-locutrices courants dans le monde (OIF 2022)
- Officielle dans ~30 pays
- Répartition géographique étendue:
  - discontinue
  - contact avec d'autres langues

  → variation dialectale

L’exploration du français

- 2 principaux types de données vocales
  - Regroupement d’enregistrements audio d’autres sources comme
    - les livres audio
      - ex. LibriSpeech pour l’anglais (Panayotov et al. 2015, www.openslr.org/12)
    - la télévision ou la radio
      - ex. OSEO Quaero (www.quaero.org)
    - l’enregistrement des débats publics
      - ex. AssNat (ANF 2011, ANQ 2011)
  - Collections de grands corpus « faites maison »
    - en laboratoire
      - ex. NCCFr pour le français (Torreira et al. 2010)
    - sur le terrain
      - ex. PFC (Durand et al. 2002)

L’exploration de la variation dialectale: défis

- Logistiques
  - Recrutement des participant·e·s, lieu d’enregistrement…
  - Grands corpus multilingues / multidialectes
- Ressources
  - Financements (personnel, matériel, transports)
  - Temps (de récolte, de traitement)
  - Energie (humaine, coût carbone)

L’exploration de la variation dialectale: solution ?

- Données participatives
  - Données (vocales) issues d’initiatives citoyennes
    - ex. Common Voice (Ardila et al. 2020): http://commonvoice.mozilla.org
  - administrées et enregistrées par des volontaires = moins coûteux
  - généralement open-source = pas de problème d’accessibilité

→ Ce genre de données est-il fiable ?

→ Permet-il d’explorer des questions de variation ?

→ Jusqu’où peut-on aller avec ?

Plan

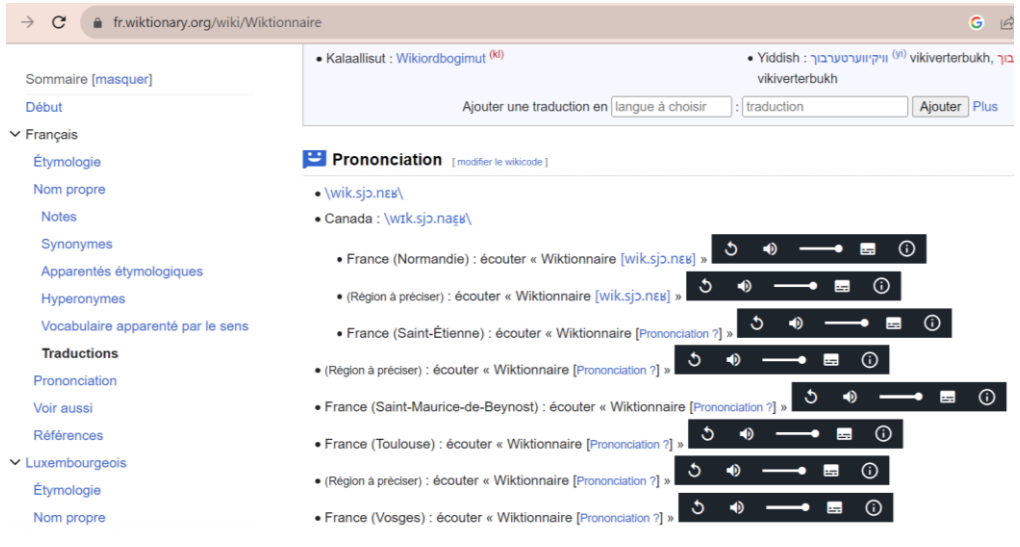
1. Lingua Libre
2. LiLi-Fr est-il un « bon » jeu de données ?
3. LiLi-Fr à l’épreuve: L’opposition /a~ɑ/ dans les français du monde
4. Bilan : Potentiel, limites et conseils d’utilisation

Lingua Libre

Données et méthode

Lingua Libre, un projet Wikimédia France

- Médiathèque linguistique participative développée par Wikimédia France
  - www.lingualibre.org
  - visée encyclopédique (connaissance) et patrimoniale (conservation + valorisation des langues)
  - utilisée notamment pour nourrir Wikipédia et le wiktionnaire



- Version beta lancée en 2015, version alpha en 2018
- Utilisé une seule fois dans un travail académique (Marjou, 2021)

Lingua Libre en octobre 2024

- 1 287 228 enregistrements

Carte générable en ligne en temps réel. Merci à Lucas Pregaldiny!

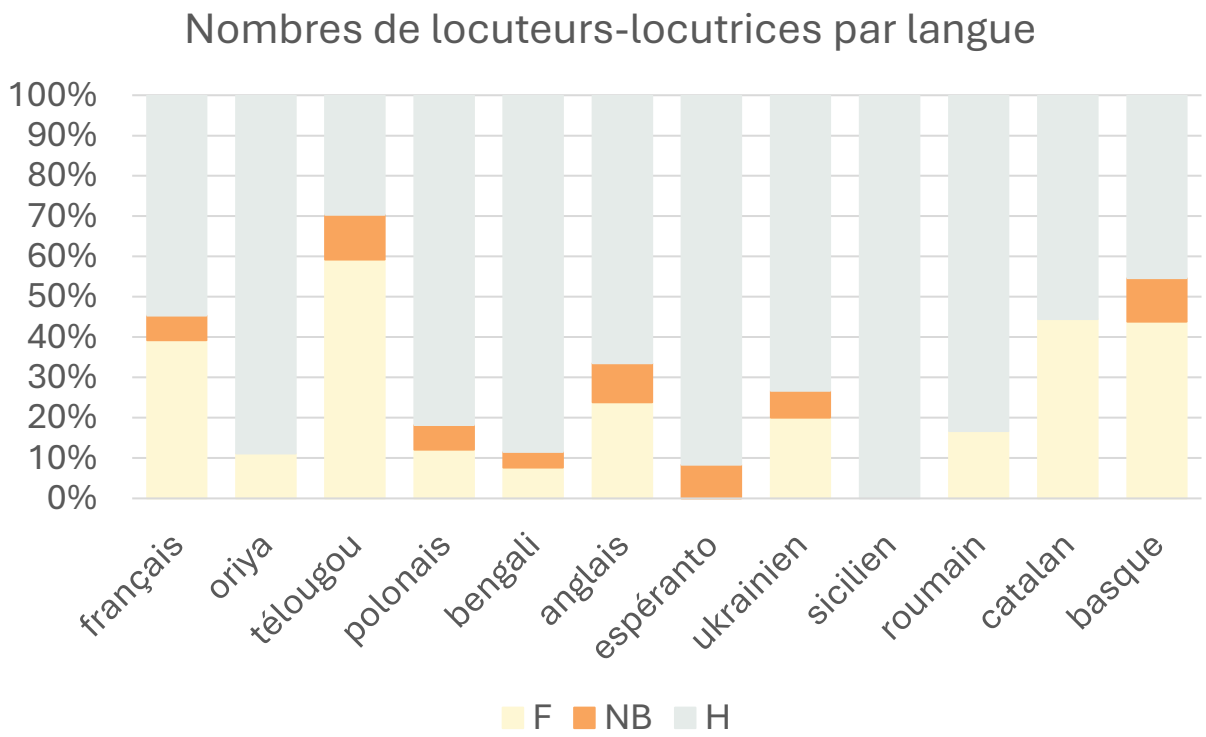
Lingua Libre en octobre 2024

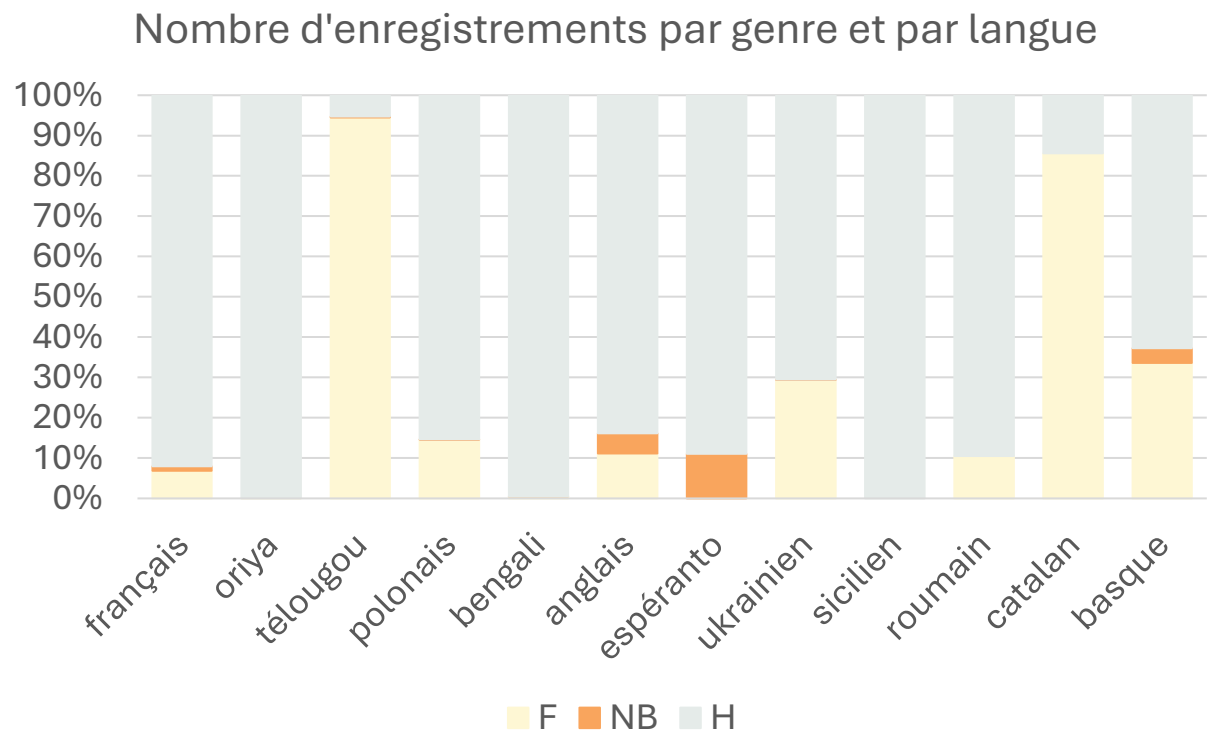
- 257 langues, dont 12 avec plus de 20.000 enregistrements
- Nombreuses familles de langues: **indo-européennes** (romanes, indo-iraniennes, slaves, germaniques, celtiques, helléniques…), dravidiennes, turciques, afro-asiatiques (sémitiques), sino-tibétaines (chinoises)… et même inventées!
- Nombreuses langues « sous-dotées », rares, voire en danger…

Langue	Nombre d'enregistrements	Langue	Nombre d'enregistrements	Langue	Nombre d'enregistrements	Langue	Nombre d'enregistrements
français	386 640	catalan	22 888	portugais	7 599	turc	4 412
oriya	131 377	basque	20 254	azéri	6 945	norvégien	4 365
télougou	98 536	allemand	18 602	arabe	6 887	pendjabi	4 362
polonais	94 645	marathi	16 702	cantonais	6 683	persan	4 116
bengali	68 660	espagnol	16 451	bulgare	6 027	kurmandji	3 857
anglais	62 703	occitan	14 173	occitan languedocien	5 327	asturien	3 686
espéranto	35 298	russe	13 915	tamoul	5 183	breton	3 452
ukrainien	27 208	malayalam	12 972	odia du Baleswar	5 096	arabe levantin méridional	3 314
sicilien	23 249	suédois	8 888	occitan gascon	4 915	hébreu	3 114
roumain	22 981	italien	8 412	occitan sifflé d'Aas	4 672	afrikaans	2 943

Lingua Libre en octobre 2024

- 2197 locuteurs et locutrices
  - Femmes = **34,49%** (μ=31,5 ; σ=68,11, +2,91%), non-binaires = **7,21%** (μ=6,58 ; σ=11,38, +2,11%)
  - Enregistrés par des femmes = **17,88%** (-6,11%), par non-binaires = **1,25%** (-0,05%)





Lingua Libre pour le français

LiLi-Fr est-il un “bon” jeu de données ?

LiLi-FR est-il un « bon » jeu de données?

- Pour permettre la reproduction des études, les données doivent être accessibles à tout le monde (Perkel 2023) → **accessibilité**
- Pour refléter au mieux la langue, les données doivent représenter différents types de parole → **variété des usages**
- Pour être utilisables, les données doivent être accompagnées de métadonnées (Perkel 2023) → **exploitabilité**
- Plus les phénomènes sont fins et multifactoriels, plus il faut de données (Coleman et al. 2016) → **quantité**
- Pour les études typologiques, les données doivent être multilingues (Salesky et al. 2020) → **comparabilité** (trans-linguistique)
- Pour éviter les biais, il faut que les données soient équilibrées (Benzeghiba et al. 2007) → **représentativité**
- Pour constituer un corpus, les données vocales doivent (pouvoir) être transcrites / alignées / annotées → **augmentabilité**
- Pour être bien traitées, les données vocales doivent être « propres » (Hutin et al. 2022ab) → **qualité**

Lingua Libre, un outil facile d'utilisation

- Processus d'enregistrement:
  - Connexion (compte Wikimédia nécessaire)
  - « Enregistrer » → Test du micro
  - Création du profil
  - Choix des mots à enregistrer (et dans quelle langue): création de liste ou choix d'une liste locale
  - Lancer l'enregistrement: les fichiers sons se segmentent tout seuls!
  - Vérification des fichiers audio.
  - Versement des fichiers validés dans Wikimedia Commons

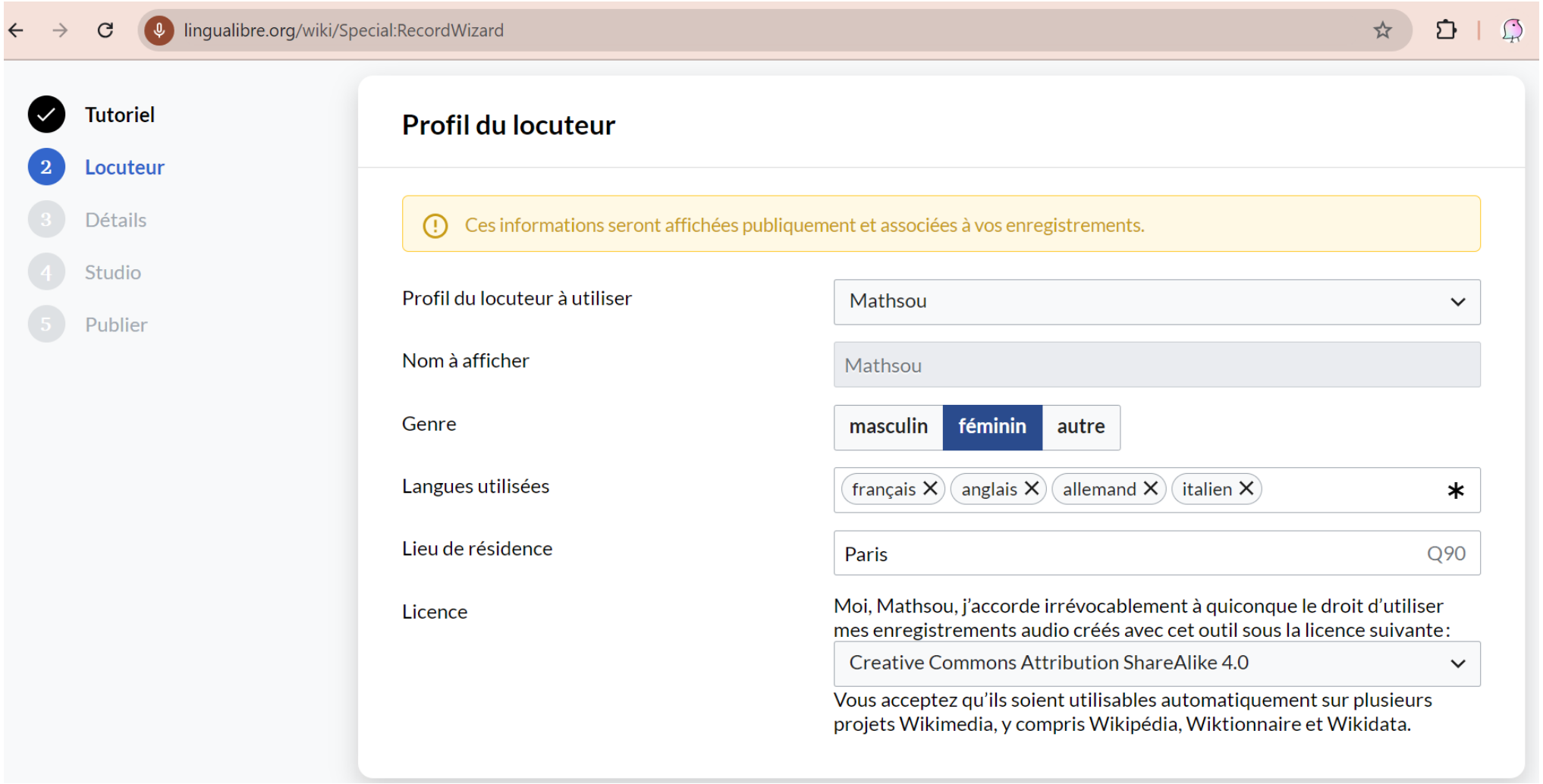
LiLi-Fr est-il un « bon » jeu de données?

→ Accessibilité ✓

→ Variété des usages X

→ Exploitabilité ?

Les métadonnées de Lingua Libre



LiLi-Fr est-il un « bon » jeu de données?

→ Accessibilité ✓

→ Variété des usages X

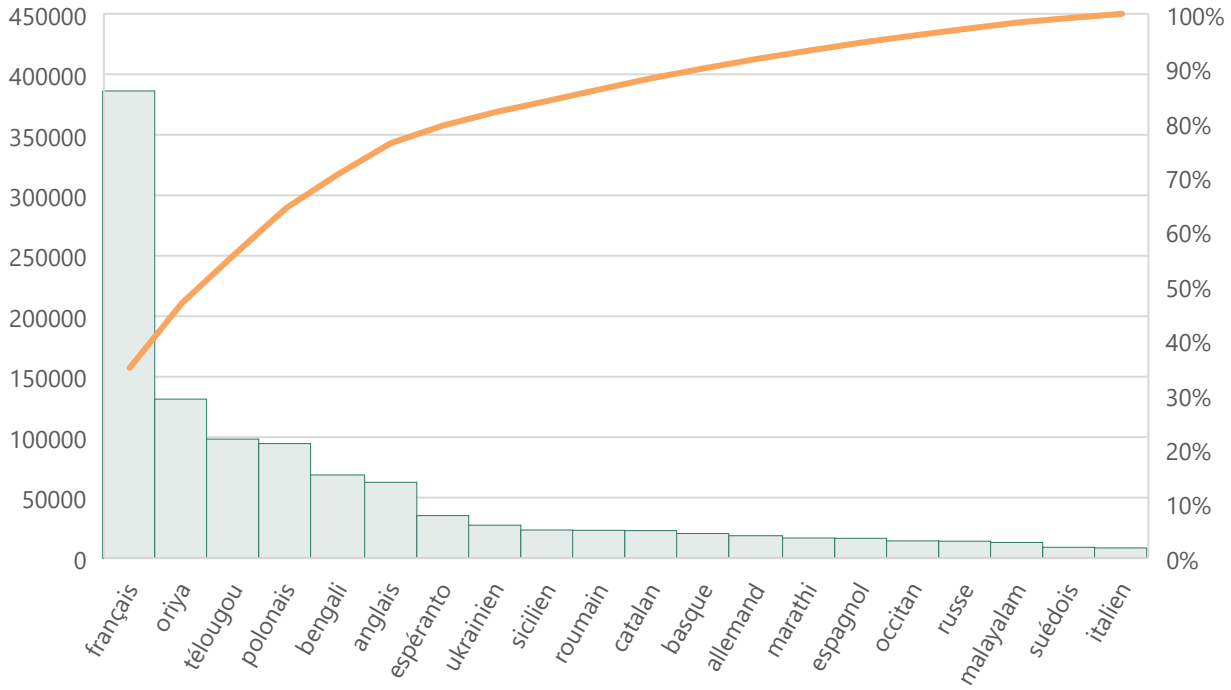
→ Exploitabilité ✓

→ Quantité ?

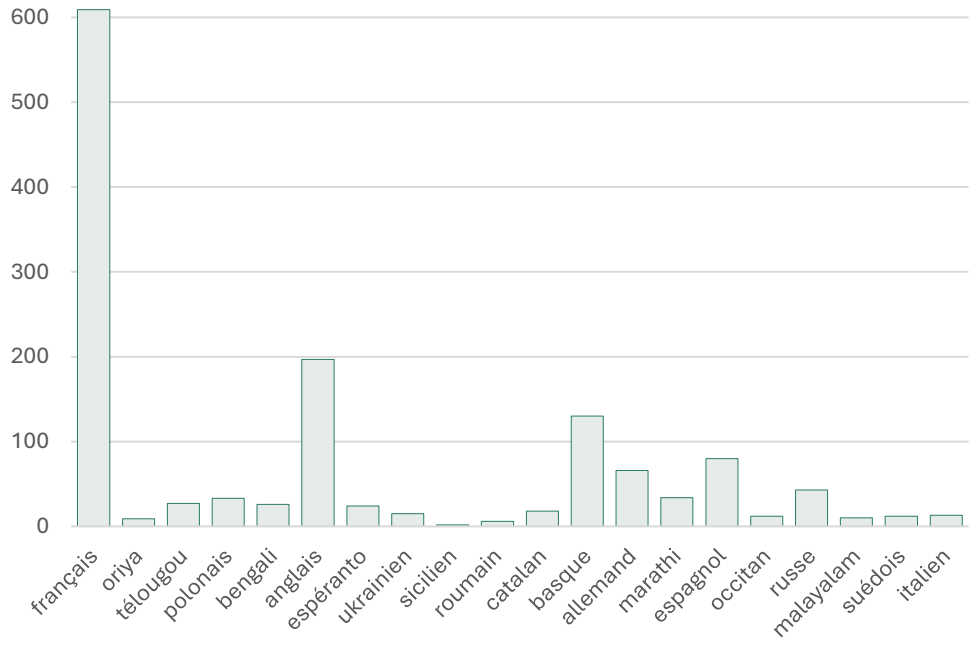
Les données de LiLi-Fr

- 386 640 enregistrements par 612 locuteurs-locutrices

Diagramme de Pareto du nombre d'enregistrements dans les 20 premières langues de LiLi



Nombre de locuteurs et locutrices dans les 20 premières langues de LiLi



LiLi-Fr est-il un « bon » jeu de données?

→ Accessibilité ✓

→ Variété des usages X

→ Exploitabilité ✓

→ Quantité ✓

→ Comparabilité ?

LiLi-Fr à travers le monde

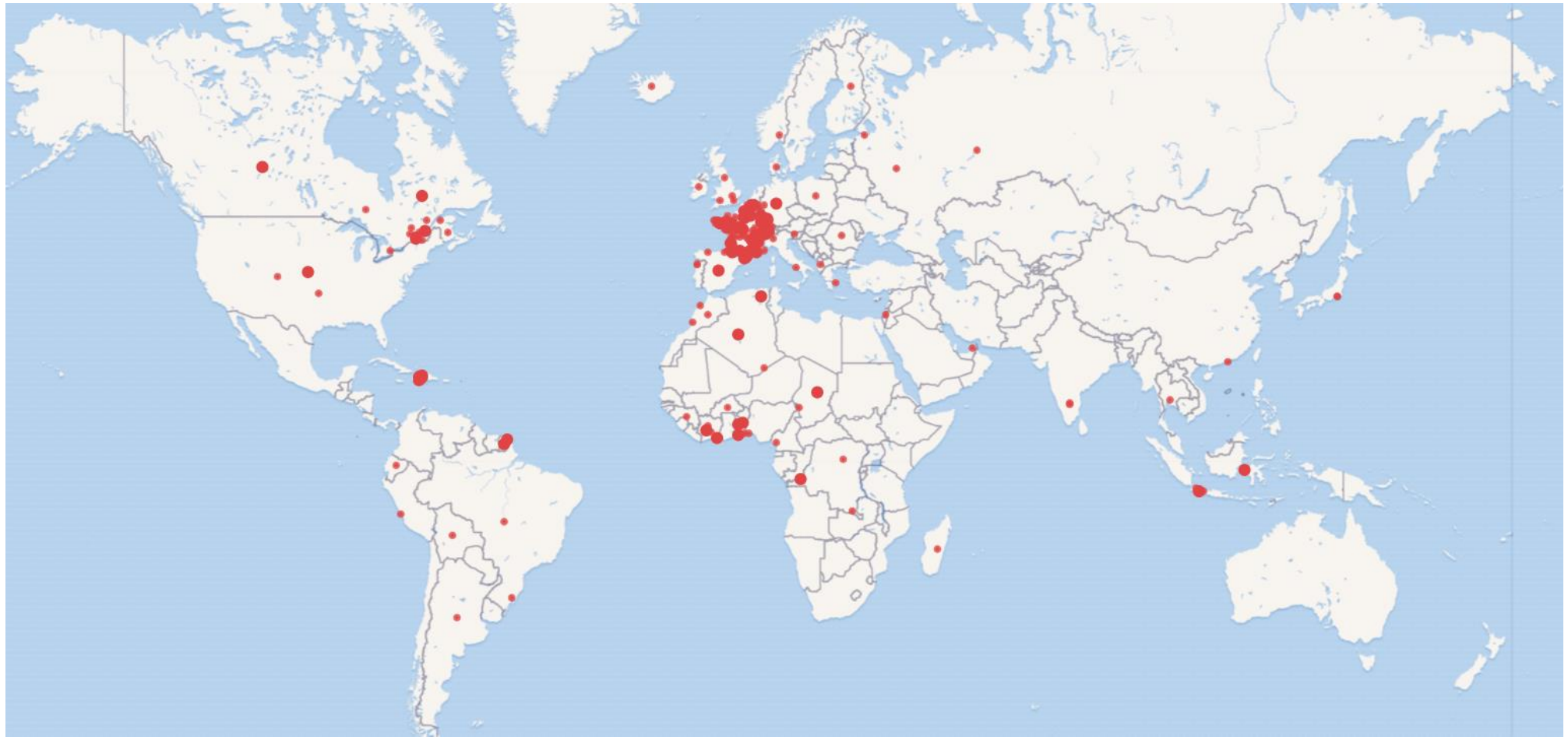


Figure. Répartition géographique des francophones dans Lingua Libre en octobre 2024.
Source: https://w.wiki/8CDt (un grand merci à Lucas Prégaldiny!)

LiLi-Fr est-il un « bon » jeu de données?

→ Accessibilité ✓

→ Variété des usages X

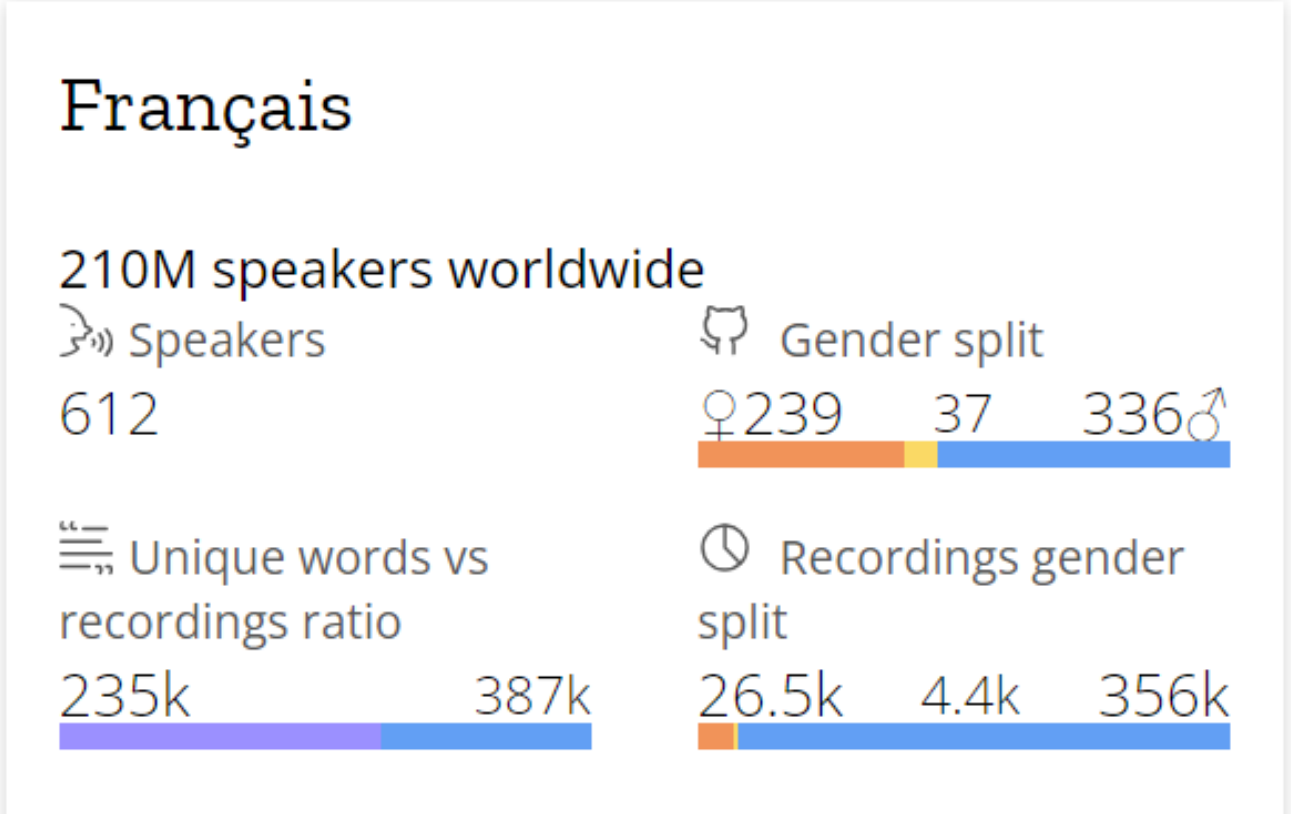
→ Exploitabilité ✓

→ Quantité ✓

→ Comparabilité ✓

→ Représentativité ?

Répartition des profils dans LiLi-Fr



LiLi-Fr est-il un « bon » jeu de données?

→ Accessibilité ✓

→ Variété des usages X

→ Exploitabilité ✓

→ Quantité ✓

→ Comparabilité ✓

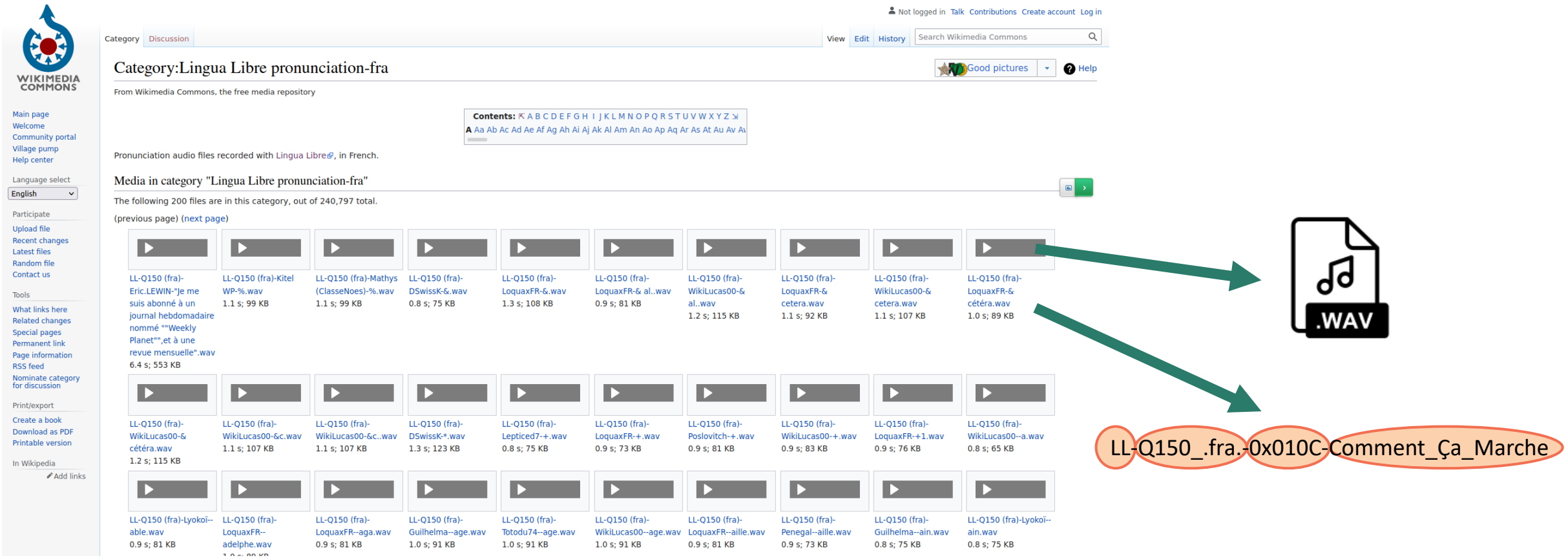
→ Représentativité ≈

→ Augmentabilité ?

Procédure d’enrichissement

1. Extraire les fichiers audios
2. Générer automatiquement les transcriptions
3. Générer automatiquement les alignements

Etape 1: Extraire les données audio



Etape 2 : Générer les transcriptions

LiLi_transcriptions

Fichier Modifier Affichage Insérer Exécution Outils Aide Dernière modification effectuée le 9 octobre

+ Code + Texte

Code for the extraction of transcriptions from file names

Mathilde Hutin

Sept, 12th 2024

```
[ ] from google.colab import files # For future save
```

```
[ ] import os, sys
    os.listdir("C:/Users/mhutin/Documents/LiLi/(I)PFC/Données") # to list file names. You may have to change the path to the files.
    path = "C:/Users/mhutin/Documents/LiLi/(I)PFC/Données"
    file_list = os.listdir(path)
    print(file_list)
    print(len(file_list))
```

```
['File_LL-Q150 (fra)-val (Eavq)-bâton.wav', 'File_LL-Q150 (fra)-LoquaxFR-bâiller.wav', 'File_LL-Q150 (fra)-Mona (Mathsou)-halle.wav', 'File_LL-Q150
337
```

```
[ ] # Open each txt file automatically
    for file_name_complete in file_list:
      file_name = file_name_complete[:-4] # to delete the ".wav" at the end of the string
      print(file_name)
      file_name_info = file_name.split("-") # to turn file name into list of infos
      print(file_name_info)
      transcript = file_name_info[-1] # to extract the transcription (last item on the list)
      print(transcript)
      save_path = "C:/Users/mhutin/Documents/LiLi/(I)PFC/Données"
      new_file_name = os.path.join(save_path, file_name +".txt") # to create the file name with .txt extension
      print(new_file_name)
      file = open(new_file_name,'w',encoding="UTF-8") # to open a new txt file. UTF8 encoding is necessary for the emuR package to work on R later.
      file.write(transcript) # to write the transcription in it
      file.close() # to close it
      #files.download(new_file_name)
```

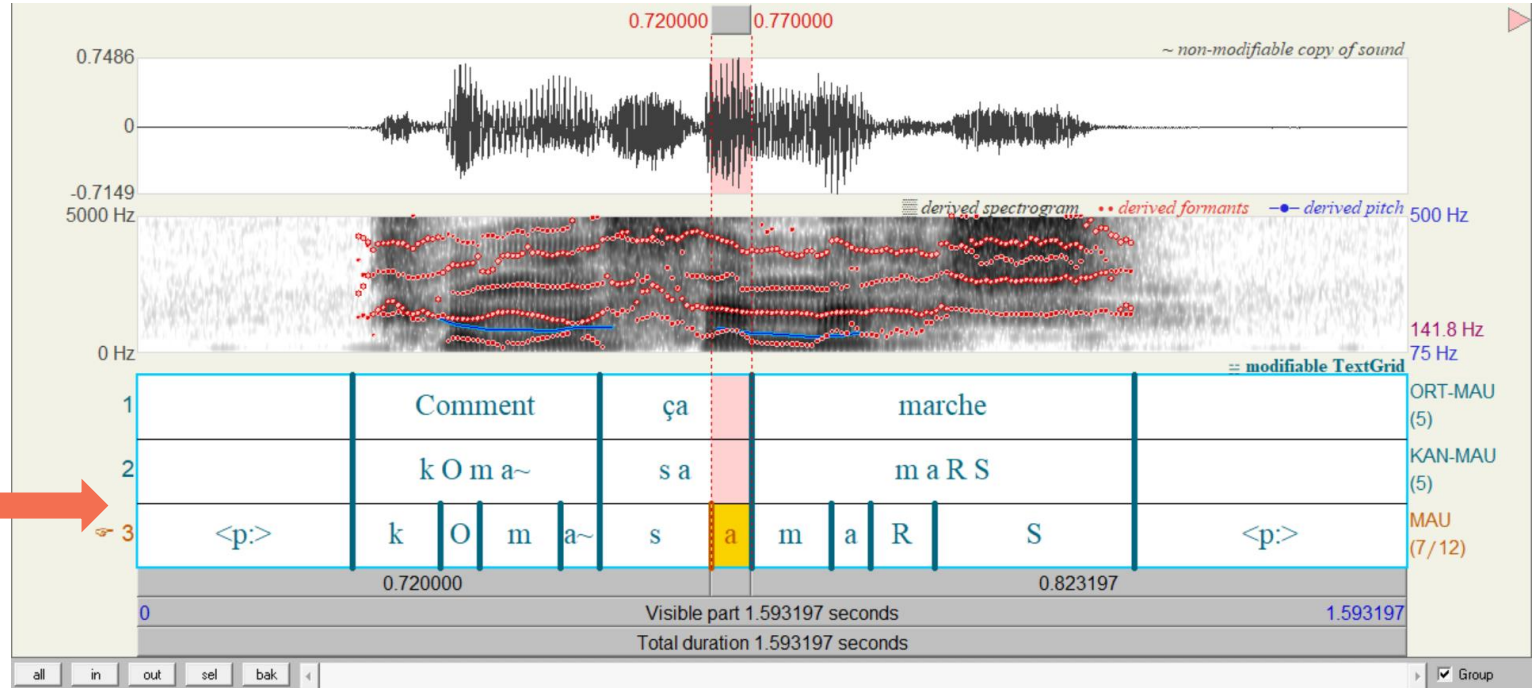
Afficher la sortie masquée

Etape 3 : Aligner audio et transcription

1. Déduire la transcription phonétique grâce à un convertisseur de graphème à phonème
2. Aligner la transcription avec l'audio.

Comment ça marche

kOma~ sa maRS



Etape 3 : Aligner audio et transcription

- WebMAUS (Kisler et al. 2012):
  - Version en ligne de MAUS (Schiel 1997, 1999, 2004)
  - Grapheme-to-phoneme converter intégré (BALLOON par Uwe Reichel)
- Avantages:
  - Nombreuses fonctionnalités
  - Éthique (ouvert, protection des données…)
  - Communauté réactive (notamment F. Schiel lui-même)
  - 43 langues ou variétés de langues (mais seulement 1 variété de français)
  - Alignement relativement fiable:
    - L'alignement WebMAUS correspond à 95% aux alignements réalisés manuellement (Kipp et al. 1997).
    - Entre l'alignement WebMAUS et sa correction manuelle, il y a un écart de 0,01 seconde en moyenne (Hutin & Allassonnière-Tang 2023).

LiLi-Fr est-il un « bon » jeu de données?

→ Accessibilité ✓

→ Variété des usages X

→ Exploitabilité ✓

→ Quantité ✓

→ Comparabilité ✓

→ Représentativité ≈

→ Augmentabilité ✓

→ Qualité ?

Qualité des données : le cas du polonais

- Hutin, M. & Allassonnière-Tang, M. 2022. Crowd-sourcing for Less-resourced Languages: Lingua Libre for Polish. *Proc. of SIGUL 2022*. Marseille, France, 24-25 juin 2022, 41-47 [pdf]
- 2022 = 2e langue de Lingua Libre (~80k enregistrements)
- La plupart des corpus du polonais (sinon tous) :
  - souvent peu représentatifs d'une large proportion de la population
  - parfois limités à des sociolectes spécifiques
  - défauts techniques
  - créés dans le but assumé de développer ou d'entrainer des outils: payants / chers
- Résultats:
  - Fréquence des phonèmes très similaires
  - Valeurs de F1 et F2 moins précises (F1 et F2 moyens plus hauts)

LiLi-Fr est-il un « bon » jeu de données?

→ Accessibilité ✓

→ Variété des usages X

→ Exploitabilité ✓

→ Quantité ✓

→ Comparabilité ✓

→ Représentativité ≈

→ Augmentabilité ✓

→ Qualité ✓

LiLi-Fr à l'épreuve :
L'opposition /a~ɑ/ dans les français du monde

Hutin, Mathilde & Allassonnière-Tang, Marc. 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. https://doi.org/10.1017/S0959269523000200

/a/ *vs* /ɑ/ : une situation complexe

- /a/ = voyelle traditionnellement notée comme ouverte antérieure voire centrale
- /ɑ/ = voyelle traditionnellement notée comme ouverte postérieure

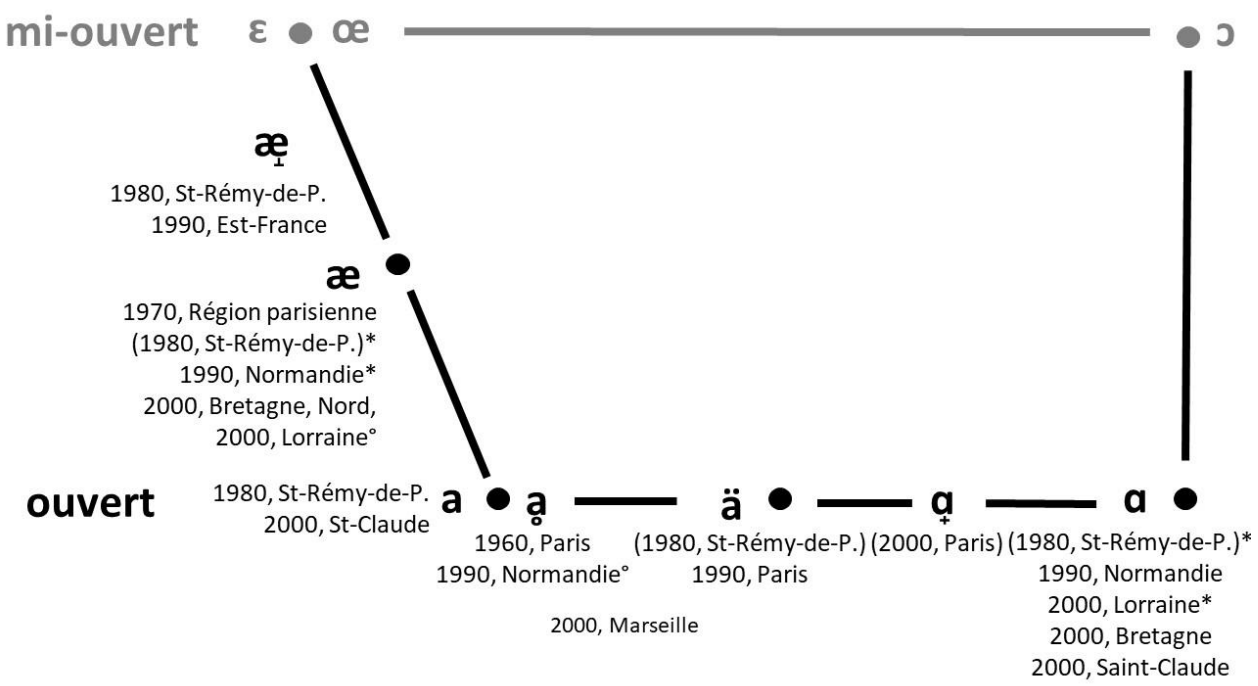


Figure. Réalisations possibles de /a/ et /ɑ/ en français de France dans la deuxième moitié du XXe siècle : dans les années 1960 à Paris (Martinet 1969), 1970 en région parisienne (François 1974, Lennig 1978, Mettas 1979), 1980 à Saint-Rémy-de-Provence dans les Bouches-du-Rhône (Hilt1986), 1990 à Paris (Fougeron and Smith 1999), en Normandie (Hauchecorne and Ball1997) et dans l'est de la France (Armstrong1993), 2000 dans le Nord et l'Alsace (Woehrling2009), en Lorraine et en Bretagne (Boughton2005), à Paris (Hansen et al. 2012), à Saint-Claude dans le Haut-Jura (Arnaud2006) et à Marseille (Coquillon et al. 2012). Les voyelles orales mi-ouvertes du français sont indiquées en gris à titre de repère. Les encarts spécifient pour chaque réalisation la variété où elle a été observée : les variantes entre parenthèses sont des variantes rares, les variantes avec * ont été observées essentiellement en syllabe fermée et les variantes avec ° ont été observées essentiellement après /w/.

/a/ *vs* /ɑ/ en (très) simplifié

- Afrique : Opposition neutralisée
- Belgique : Opposition maintenue mais essentiellement sur la durée
- Canada : Opposition maintenue (postériorité et/ou longueur/diphtongaison)
- France :
  - Centre : Opposition majoritairement neutralisée
  - Nord-Est : Opposition neutralisée, éventuellement maintenue en syllabe fermée
  - Nord-Ouest : Opposition maintenue dans toutes les positions
  - Sud-Est : Opposition généralement neutralisée
  - Sud-Ouest : Opposition généralement neutralisée
- Suisse : Opposition maintenue en syllabe fermée, mais beaucoup de variation en syllabe ouverte selon les cantons

Sélection des paires minimales

En syllabe fermée finale	
/a/	/ɑ/
Anne	âne, ânes
bal	Bâle
mal	mâle, mâles
pack, packs	Pâques
pal	pâle, pâles
patte, pattes	pâte, pâtes
tache, taches	tâche, tâches

En syllabe ouverte finale	
/a/	/ɑ/
bas	bât
ma	mât, mâts
rat, rats	ras

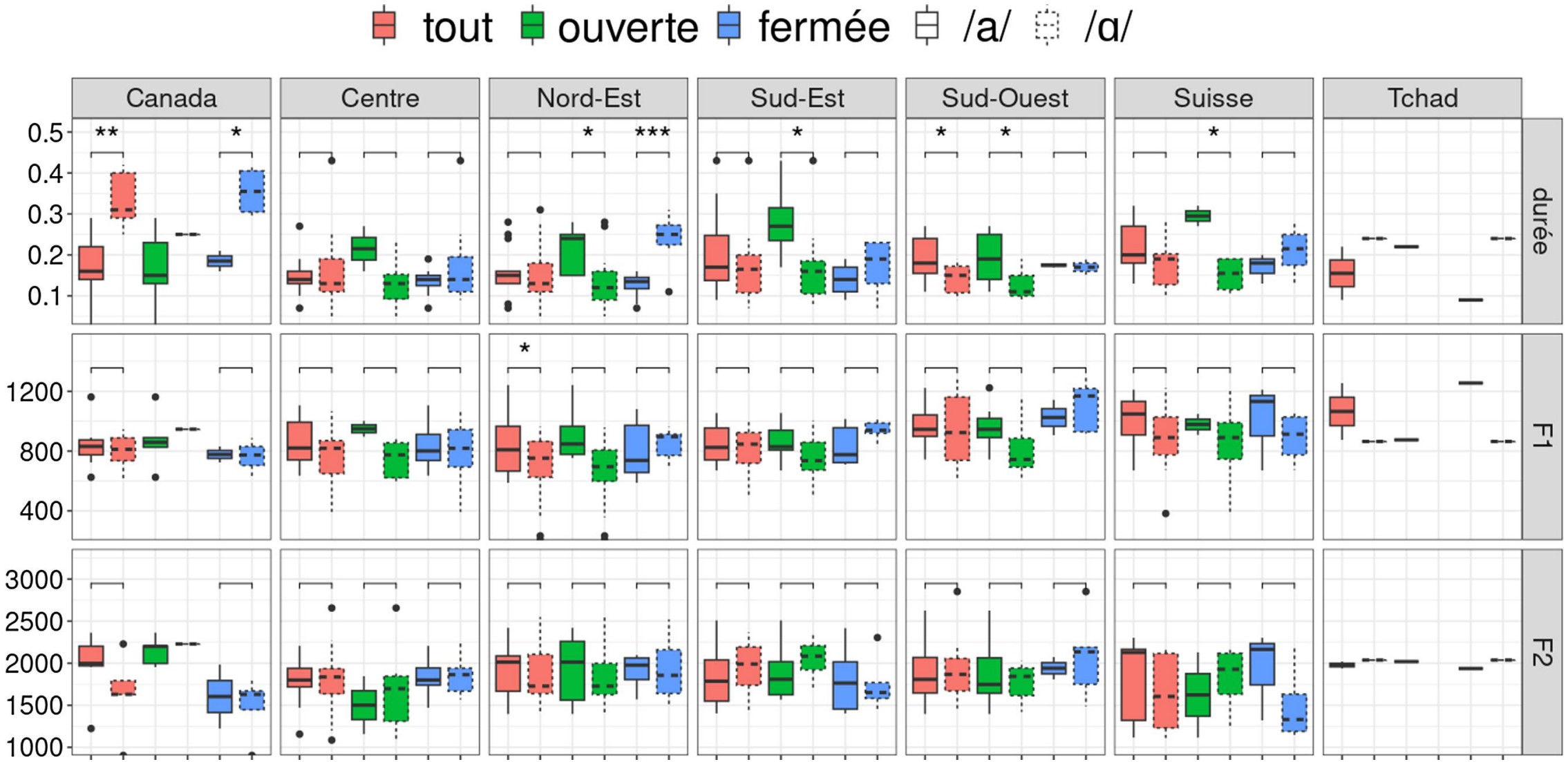
En syllabe ouverte non-finale	
battons	bâton, bâtons
bailler, bayer (+conj.)	bâiller (+conj.)

Filtrage des données

- 38 personnes dont la région est renseignée assez précisément
  - 9 femmes, 28 hommes, 1 non-déclarée
- 163 enregistrements
  - 67 /a/ et 96 /ɑ/
  - (sur « seulement » 239 457 enregistrements produits par 255 locuteurs)
- 26 localités différentes sur 7 zones

	/a/	/ɑ/	**Total**
Canada	7	5	**12**
Centre	13	17	**30**
Nord-Est	13	37	**50**
Sud-Est	16	16	**32**
Sud-Ouest	11	12	**23**
Suisse	5	8	**13**
Tchad	2	1	**3**
Total	**67**	**96**	**163**

Résultats


tout
ouverte
fermée
/a/
/ɑ/
Canada
Centre
Nord-Est
Sud-Est
Sud-Ouest
Suisse
Tchad
durée
F1
F2
0.5
0.4
0.3
0.2
0.1
1200
800
400
3000
2500
2000
1500
1000

LiLi-Fr pour explorer la variation dialectale?

Bilan: potentiel, limites et conseils

Potentiel

- Moindres coûts
- Beaucoup de données… and counting!
- Analyses phonético-phonologiques possibles
  - Mais il faut ajuster les phénomènes à observer…

Limites

- Analyses acoustiques moins précises ?
- Uniquement de la parole lue… pour le moment !
- Données et métadonnées (à moitié) imposées

Que faire ?

- Ajuster la finesse du phénomène à la quantité de données
  - Sinon, ajouter ses propres enregistrements…
- Utiliser LiLi comme plateforme d'enregistrement et de stockage
  - Valorisation des données de la recherche en bonus!
  - Enregistrer avec un bon micro
  - Expliquer pourquoi les métadonnées sont importantes…
- Usage exploratoire
  - Explorer une langue / variété peu connue (ex. pour préparer son terrain)
- Usage expérimental
  - Utiliser les fichiers-sons pour des tests perceptifs (ex. dialectologie perceptuelle, cf. Preston 1989)

Un très grand merci à…

Lucas Pregaldiny / WikiLucas
Administrateur de Lingua Libre
Paris, France

Marc Allassonnière-Tang
Chargé de recherche
Centre National pour la Recherche Scientifique (CNRS)
Laboratoire Eco-Anthropologie (EA), UMR 7206
Musée National d'Histoire Naturelle (MNHN)
Paris, France

Lucas Lévêque / Lyokoï
Co-fondateur de Lingua Libre
Administrateur du Wiktionnaire
Fondateur et PDG du Lug numérique
Lyon, France

Merci de votre attention!

Questions ? Remarques ?

mathilde.hutin@uclouvain.be

Références

ANF. 2011. Assemblée nationale de France. http://www.assemblee-nationale.fr/index.asp. Website of Assemblée nationale de France.

ANQ. 2011. Assemblée nationale du Québec. http://www.assnat.qc.ca/fr/index.html. Website of Assemblée nationale du Québec.

Ardila, R., et al. 2020. *Common voice: A massively-multilingual speech corpus*. arXiv arXiv:1912.06670.

Armstrong, N. 1993. *A study of phonological variation in French secondary school pupils*. Unpublished PhD thesis, University of Newcastle upon Tyne.

Arnaud, V. 2006. *La dimension variationniste du français en usage à Saint-Claude (Haut-Jura) : une étude acoustique des voyelles orales des « gens d'en haut »*. Thèse de doctorat, Université Laval (Québec) et Université de Franche-Comté (France).

Benzeghiba, M., R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens. 2007. Automatic speech recognition and speech variability: A review, *Speech Communication* 49: 10–11, 763-786, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2007.02.006.

Boughton, Z. 2005. Accent leveling and accent localisation in northern French: Comparing Nancy and Rennes. *Journal of French Language Studies*, 15, 235–256. https://doi.org/10.1017/S0959269505002140

Coleman, J., M. E.L. Renwick & R.A.M. Temple. 2016. Probabilistic underspecification in nasal place assimilation, *Phonology*, 33(3), 425-458.

Coquillon, A. & Turcsan, G.. 2012. An overview of the phonological and phonetic properties of Southern French Data from two Marseille surveys. *Phonological Variation in French: Illustrations from Three Continents*, edited by Gess, R., Lyche, C. & Meisenburg, T., John Benjamins Publishing Company. 105–127.

Durand, J., Laks, B. & Lyche, C. 2002. La phonologie du français contemporain: usages, variétés et structure. In Pusch, C. & Raible, W. (eds), *Romanistische Korpuslinguistik – Korpora und gesprochene Sprache/Romance Corpus Linguistics – Corpora and Spoken Language*. Tübingen : Gunter Narr Verlag, 93–106. PDF (Durand/Laks/Lyche 2002)

Fougeron, C. & Smith, C. L. 1999. French. *Handbook of the International Phonetic Association*, 78–81. Cambridge: Cambridge University Press.

François, D. 1974. Français parlé : analyse des unités phonétiques et significatives d'un corpus recueilli dans la région parisienne. Paris, *S.E.L.A.F.*

Garofolo, J., Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, Victor Zue. 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium. https://doi.org/10.35111/17gk-bn40

Hansen, A. B. 2012. A study of young Parisian speech: Some trends in pronunciation. *Phonological Variation in French : Illustrations from Three Continents*, edited by Gess, R. Lyche, C. & Meisenburg, T., John Benjamins Publishing Company. 151–172.

Hauchecorne, F. & Ball, R. 1997. L'accent du Havre : un exemple de mythe linguistique. *Langage & société* 82, 5–25.

Hilt, A. 1986. Une analyse auditive sur les voyelles à double timbre du français parlé à Saint-Rémy-de-Provence. *Revue Romane* 21, 2.

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~ɑ/. *Journal of French Language Studies*, 1-24. doi: https://doi.org/10.1017/S0959269523000200

Hutin, M. & Allassonnière-Tang, M. 2022a. Crowd-sourcing for Less-resourced Languages: Lingua Libre for Polish. *Proceedings of SIGUL 2022*. Marseille, France, 24-25 juin 2022, pp. 41-47.

Références

Hutin, M. & Allassonnière-Tang, M. 2022c. 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., Weng, C., Adda-Decker, M. & Lamel, L. 2022. La liaison facultative en français : étude de grands corpus combinant approche automatique relâchée et jugement perceptif. *Actes du 8e Congrès Mondial de Linguistique Française – CMLF 2022*, 4-8 juillet. doi: https://doi.org/10.1051/shsconf/202213810004

Hutin, M., Weng, C., Adda-Decker, M., Vasilescu, I. & Lamel, L.. 2022. Disfluences et erreurs d'alignement au niveau du phonème : le cas des consonnes de liaison en français. *Actes des XXXIVe Journées d'Étude sur la Parole – JEP 2022*. Noirmoutier, France, 13-17 juin 2022, pp. 452-461. doi: https://doi.org/10.21437/JEP.2022-48

Kipp, A., Wesenick, M-B. & Schiel, F. 1997. Pronunciation modeling applied to automatic segmentation of spontaneous speech. *Proc. of Eurospeech*, 1023–1026.

Kisler T, Schiel F, Sloetjes H. 2012. Signal processing via web services: the use case WebMAUS. *Proceedings Digital Humanities 2012*, Hamburg, Germany (pp. 30-34).

Lennig, M. 1978. *Acoustic Measurement of Linguistic Change: The Modern Paris Vowel System*. Ph.D. dissertation, University of Pennsylvania.

Marjou, X. 2021. OTEANN: Estimating the Transparency of Orthographies with an Artificial Neural Network. Stroudsburg: Association for Computational Linguistics.

Martinet, A. 1969. « C'est jeuli, le Mareuc ! ». *Le français sans fard*. Paris, Presses Universitaires de France.

Mettas, O. 1979. *La prononciation parisienne: Aspects phonique d'un sociolecte parisien (du Faubourg Saint-Germain à la Muette)*. Paris: Conseil International de la Langue Française et Laboratoire des Langues et Civilisations à Tradition Orale du CNRS.

Organisation internationale de la francophonie. 2022. *La langue française dans le monde. Synthèse 2022*. Gallimard. https://www.francophonie.org/sites/default/files/2023-03/Rapport-La-langue-francaise-dans-le-monde_VF-2022.pdf

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 5206–5210.

Perkel J. M. 2023. How to make your scientific data accessible, discoverable and useful. *Nature*, 618(7967), 1098–1099. https://doi.org/10.1038/d41586-023-01929-7

Preston, D. R. 1989. *Perceptual dialectology*. Berlin, New York: De Gruyter Mouton.

Salesky, E., Chodroff, E., Pimentel, T., Wiesner, M., Cotterell, R., Black, A. W. & Eisner, J. 2020. A corpus for large-scale phonetic typology. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 5–10. Stroudsburg: Association for Computational Linguistics, pp. 4526–46.

Schiel, F. 1997. Probabilistic analysis of pronunciation with MAUS; in: *The ELRA Newsletter*, December 1997, 6-9.

Schiel F. 1999. Automatic Phonetic Transcription of Non-Prompted Speech, *Proc. of the ICPhS 1999*. San Francisco, August 1999. 607-610.

Schiel, F. 2004. MAUS Goes Iterative. *Proc. of the IV. International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 1015-1018.

Torreira, F., Adda-Decker, M. & Ernestus, M. 2010. The Nijmegen Corpus of Casual French. *Speech Communication*, Elsevier: North-Holland, 52(3), 201–212.

Woehrling, Cécile. 2009. *Accents régionaux en français : perception, analyse et modélisation à partir de grands corpus*. Informatique [cs]. Université Paris Sud – Paris XI, 2009. Français. <tel-00617248>