L'apport des données collaboratives à l'exploration linguistique : Question de typologie phonétique

Mathilde Hutin & Marc Allassonnière-Tang
LISN-CNRS (UMR 9015) & EA / MNHN (UMR 7206)

Plan

1. Introduction
    a. La recherche à l'épreuve des données
    b. L'hypothèse de la taille d'inventaire
2. Données et méthode
    a. Lingua Libre
    b. Alignement et extraction des formants
3. Résultats
    a. Hypothèse invalidée
    b. Mais…
4. Conclusion et Discussion

***Merci à Lucas Lévêque / Lyokoï pour son expertise et son soutien.**

Introduction

La recherche à l’épreuve des données

Les études **phonétiques** ont besoin:

- de données orales (en l’occurrence vocales)
- en grande quantité

Les études en **typologie** ont besoin:

- de ça, mais dans plein de langues!

Le problème des corpus

- Les **corpus** existants sont souvent:
  - non-libres de droits / difficiles d'accès => chers
  - monolingues
  - … et favorisent les langues bien dotées!

=> Utiliser des **données participatives**!

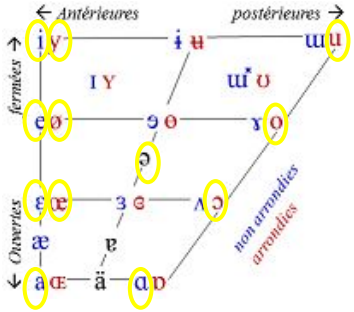
  - Ex. Common Voice de Mozilla (Ardila et al. 2020)
  - … Ou Lingua Libre de Wikimédia France!

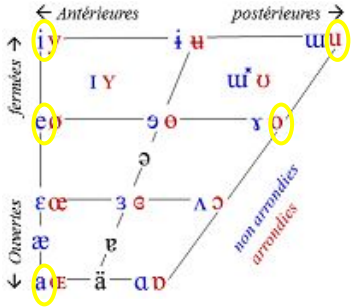
Question de typologie phonétique

Hypothèse de la taille d'inventaire (H&H, Lindblom 1990)

➔ La production des voyelles, par ex., serait tributaire du nombre de "concurrentes":

Les voyelles orales du français (fond de carte: Trapèze des voyelles API pour illustrer la représentation simplifiée de ces voyelles. Les voyelles non arrondies sont en bleu et les voyelles arrondies en rouge. Wikipedia)





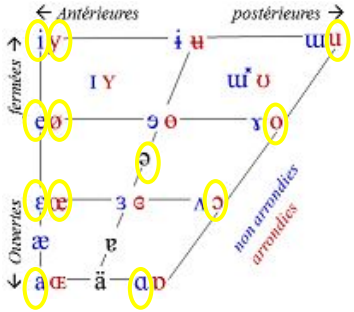
Les voyelles de l'espagnol (fond de carte: Trapèze des voyelles API pour illustrer la représentation simplifiée de ces voyelles. Les voyelles non arrondies sont en bleu et les voyelles arrondies en rouge. Wikipedia)

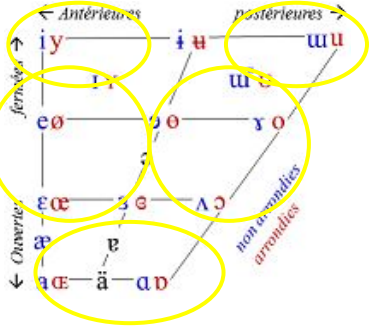
Question de typologie phonétique

Hypothèse de la taille d'inventaire (H&H, Lindblom 1990)

➔ La production des voyelles, par ex., serait tributaire du nombre de "concurrentes":

Les voyelles orales du français (fond de carte: Trapèze des voyelles API pour illustrer la représentation simplifiée de ces voyelles. Les voyelles non arrondies sont en bleu et les voyelles arrondies en rouge. Wikipedia)





Les voyelles de l'espagnol (fond de carte: Trapèze des voyelles API pour illustrer la représentation simplifiée de ces voyelles. Les voyelles non arrondies sont en bleu et les voyelles arrondies en rouge. Wikipedia)

Question de typologie phonétique

Hypothèse de la taille d'inventaire (H&H, Lindblom 1990)

➔ La production des voyelles, par ex., serait tributaire du nombre de "concurrentes":

- quelques pour (Jongman et al. 1989, Al Tamimi & Ferragne 2005, Larouche & Steffann 2018)...
- beaucoup contre (Bradlow 1995, Meunier et al. 2003, Recasens & Espinoza 2009, Lee 2012, Heeringa et al. 2015)...
- surtout dans les études de 7+ langues (Engstrand & Krull 1991, Livijn 2000, Gendrot & Adda-Decker 2007, Saleski et al. 2020).

➔ Que donne une étude sur Lingua Libre?

Données et Méthode

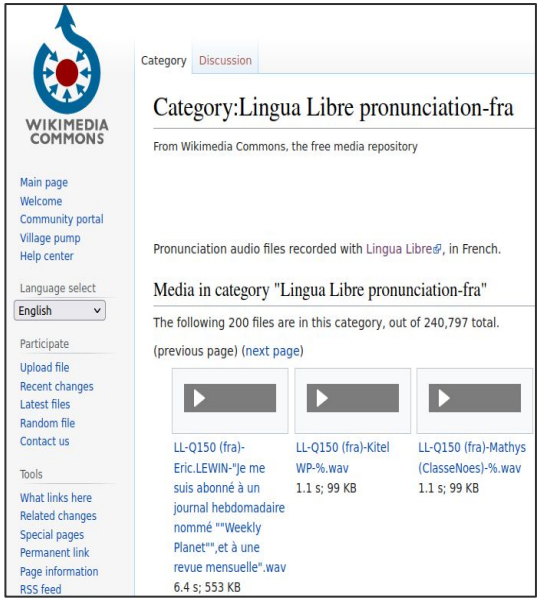
Lingua Libre

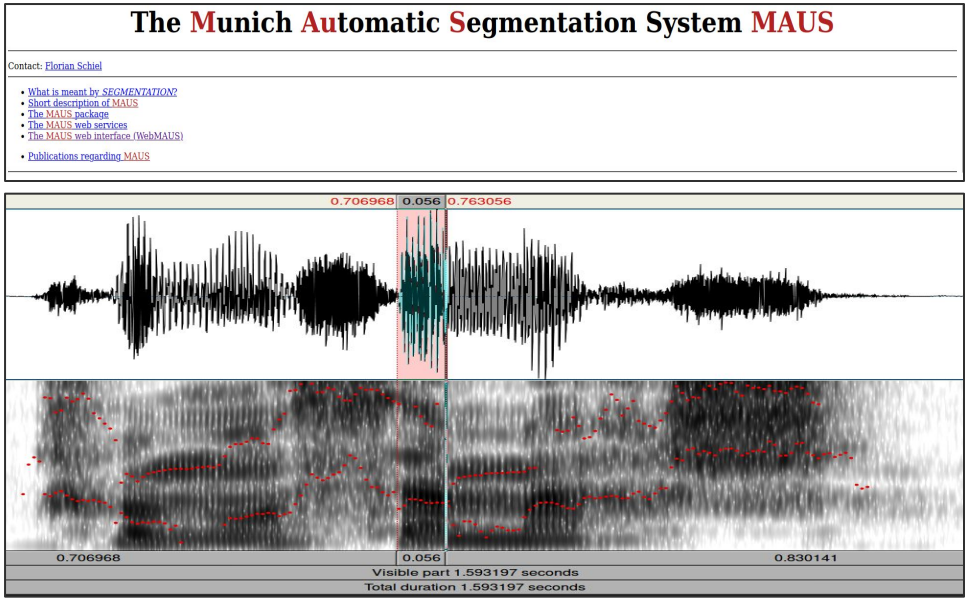
- Librairie linguistique participative de Wikimedia France
- Enregistrements de mots isolés lus à haute voix
  - ★ 687 966 entrées
  - ★ 782 locutrices et locuteurs
  - ★ 150 langues

➔ Accès à des données multilingues uniformes

Méthode

LL-Q150_.fra.-0x010C-Comment_Ça_Marche
kOma~ sa maRS





/a, i, u/ dans 10 langues indo-européennes

Table 1. Count of the vowels [a], [i], and [u] in a sample of 1000 recordings for each language in the data set. The vowel inventory refers to the number of timbers in each language.

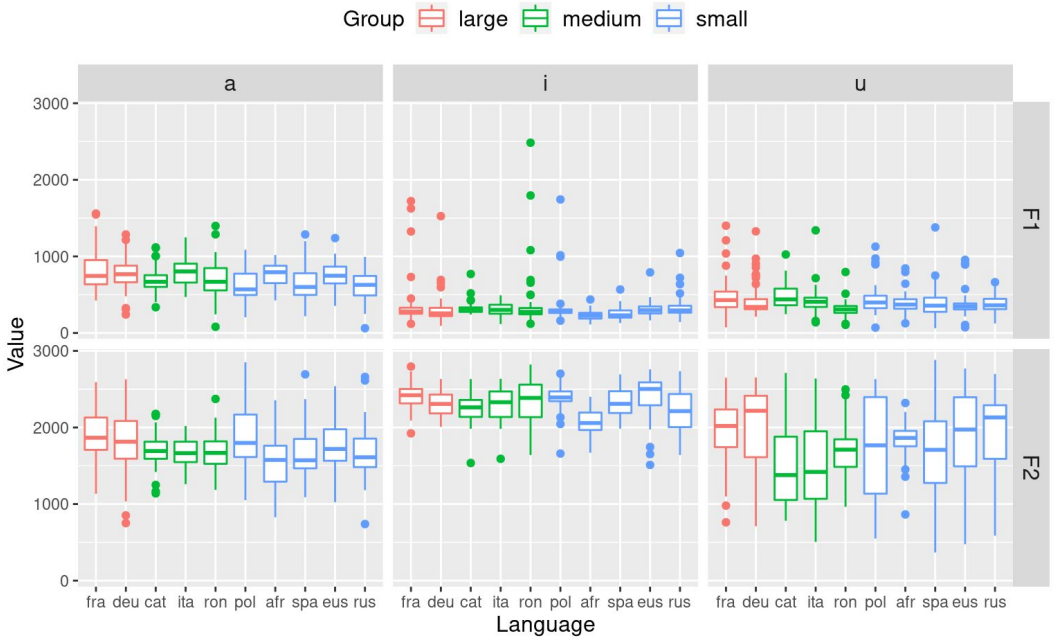
Language	iso	Vowel inventory	[a]	[i]	[u]	Speaker count
French	fra	14	408	285	56	13
German	deu	13	551	225	70	13
Catalan	cat	8	398	478	325	9
Romanian	ron	7	794	524	377	13
Italian	ita	7	969	538	104	5
Polish	pol	6	936	347	162	14
Afrikaans	afr	6	228	249	87	8
Spanish	spa	5	1087	413	139	14
Russian	rus	5	951	560	199	14
Basque	eus	5	1376	542	253	14

Source: Hutin & Allassonnière-Tang, in prep. Operation LiLi: Using crowd-sourced data and automatic alignment to investigate the phonetics and phonology of less-resourced languages. *Languages*.

*Attention: ces chiffres reflètent des analyses préliminaires et sont susceptibles de changer.

Résultats

La variation



Valeurs F1 (haut) et F2 (bas) des voyelles cardinales /a, i, u/ dans les 10 langues de l'étude. Les langues en rouge sont celles contenant beaucoup de voyelles (large), en vert celles contenant un nombre moyen de voyelles (medium) et en bleu celles contenant peu de voyelles (small).
Source: Hutin & Allassonnière-Tang (in prep).

*Attention: ces chiffres reflètent des analyses préliminaires et sont susceptibles de changer.

La variation

Table 3. The output of linear mixed models based on the output of 10 vowel samplings with 50 tokens for each vowel in each language. The abbreviations are read as follows: nb_of_timbers = number of timbers, sd = standard deviation.

Dependent variable	Predictor	Estimate	df	t value	P value
Area	nb_of_timbers	0.067	5	1.023	0.353
Area	vowel_i	-0.028	279	-0.652	0.515
Area	vowel_u	-0.009	279	-0.220	0.826
Area	Group_medium	0.468	5	1.179	0.292
Area	Group_small	0.536	5	1.063	0.336
Area	with_schwa	-0.190	5	-2.295	0.070
sd F1	nb_of_timbers	9.722	5	0.279	0.792
sd F1	vowel_i	-43.438	288	-5.569	0.000 ***
sd F1	vowel_u	-34.167	288	-4.381	0.000 ***
sd F1	Group_medium	-9.890	5	-0.047	0.965
sd F1	Group_small	-14.680	5	-0.054	0.959
sd F1	with_schwa	-14.185	5	-0.320	0.762
sd F2	nb_of_timbers	-7.188	5	-0.254	0.810
sd F2	vowel_i	-94.718	288	-9.666	0.000 ***
sd F2	vowel_u	190.762	288	19.468	0.000 ***
sd F2	Group_medium	-96.247	5	-0.558	0.601
sd F2	Group_small	-82.967	5	-0.379	0.720
sd F2	with_schwa	-88.047	5	-2.445	0.058

Source: Hutin & Allassonnière-Tang, in prep. Operation LiLi: Using crowd-sourced data and automatic alignment to investigate the phonetics and phonology of less-resourced languages. *Languages*.

*Attention: ces chiffres reflètent des analyses préliminaires et sont susceptibles de changer.

La variation

➔ Pas de corrélation avec:

- nombre de timbres vocaliques
- groupes selon la taille d'inventaire (large, medium, small)
- présence ou absence de schwa

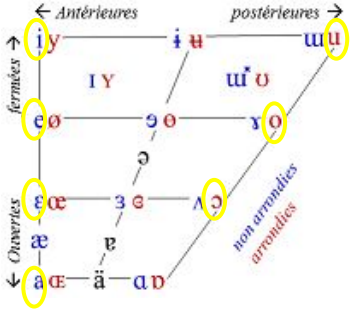
➔ L'hypothèse de la taille d'inventaire est invalidée (cf. littérature passée).

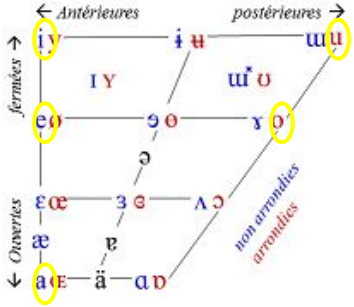
Mais…

L’hypothèse de la qualité vocalique (cf. Hutin et Allassonnière-Tang, soumis):

➔ Variation sur toutes les dimensions ou seulement une?

Les voyelles de l’italien (fond de carte: Trapèze des voyelles API pour illustrer la représentation simplifiée de ces voyelles. Les voyelles non arrondies sont en bleu et les voyelles arrondies en rouge. Wikipedia)

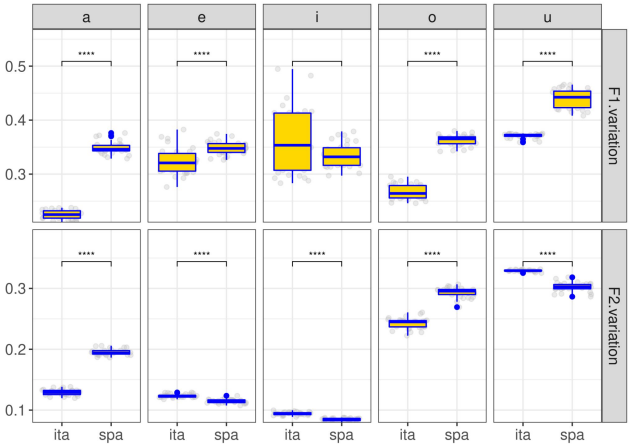




Les voyelles de l’espagnol (fond de carte: Trapèze des voyelles API pour illustrer la représentation simplifiée de ces voyelles. Les voyelles non arrondies sont en bleu et les voyelles arrondies en rouge. Wikipedia)

La qualité des voyelles, plus que leur nombre, influence la production.

Distribution du coefficient de variation pour chacun des 500 [a], [e], [i], [o] et [u] dans les données italiennes et espagnoles extraites de Lingua Libre dans chaque réplications. La significativité indique le résultat d'un test Wilcoxon avec correction Bonferroni. Source: Hutin & Allassonnière-Tang, soumis



Dep.Var	Pred	Est	t value	p value
CV F1	spa	0.05	6.97	***
CV F1	/e/	0.06	5.79	***
CV F1	/i/	0.07	6.36	***
CV F1	/o/	0.04	3.41	***
CV F1	/u/	0.12	11.19	***
CV F2	spa	0.01	3.35	**
CV F2	/e/	-0.04	-6.87	***
CV F2	/i/	-0.07	-11.64	***
CV F2	/o/	0.11	16.56	***
CV F2	/u/	0.15	23.45	***
Area	spa	212	8.981	***
Area	/e/	-88	-2.35	*
Area	/i/	-210	-5.63	***
Area	/o/	230	6.16	***
Area	/u/	196	5.25	***

Table 2: The output of linear mixed models based on the output of 10 vowel samplings with 500 tokens for each vowel in Italian and Spanish. The areas are counted as units of thousands. The abbreviations are read as follows: Pred = predictor, Est = estimate, CV = coefficient of variation, Dep.Var = Dependent variable.

Conclusion et discussion

Discussion

- Recherches futures : plus de variables? (sexe, qualité des voyelles…)
- Points forts : quantité des données, disponibilité/documentation :-)
- Remarques : difficultés(?) d'obtenir les données des locuteurs
  - LL-Pamputt-fra-lapin-garou vs. LL-Q652-ita-Yiyi-Brissago-Valtravaglia
  - téléchargement des métadonnées locuteurs (sexe, région, etc…)?
  - Système de vérification à la Common Voice?
- Tributaires des langues disponibles dans MAUS
  - Une autre méthode de segmentation? (suggestions bienvenues)
- Poste de référent scientifique / connexion recherche?

Publications prévues

- Dans des revues à comité de lecture
  - Hutin, Mathilde & Allassonnière-Tang, Marc. In prep. (abstract accepted). Operation LiLi: Using crowd-sourced data and automatic alignment to investigate the phonetics and phonology of less-resourced languages. *Languages, Advances in Phonetic Sciences: Role of Speech Corpora and Automatic Processing* (special issue)
  - Hutin, Mathilde & Allassonnière-Tang, Marc. Soumis. L'apport des données participatives pour l'étude linguistique des français du monde: le cas de l'opposition /a~ɑ/. *Langue Française*.

- Dans des actes de conférences à comité de lecture
  - Hutin, Mathilde & Allassonnière-Tang, Marc. In prep. (abstract accepted). Languages Worldwide and the World Wide Web: Crowdsourcing on the Internet to explore linguistic theories. *Proceedings of the Digital Research Data and Human Sciences Conference 2022*. Jyväskylä, Finland.
  - Hutin, Mathilde & Allassonnière-Tang, Marc. Soumis. 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)
  - Hutin, Mathilde & Allassonnière-Tang, Marc. Soumis. Crowd-sourcing for Less-resourced Languages: Lingua Libre for Polish. *Proceedings of SIGUL 2022*.

Références

Jalal-Eddin Al-Tamimi and Emmanuel Ferragne. 2005. Does vowel space size depend on language vowel inventories? Evidence from two Arabic dialects and French. *INTERSPEECH EUROSPEECH 2005*, 2465–2468, Lisbonne, Portugal.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Proceedings of LREC*.

Ann R. Bradlow. 1995. A comparative acoustic study of English and Spanish vowels. *Journal of the Acoustical Society of America*, 97:1916–1924.

Olle Engstrand and D. Krull. 1991. Effects of inventory size on the distribution of vowels in the formant space: preliminary data from seven languages. *PERILUS*, pages 15–18.

Cédric Gendrot and Martine Adda-Decker. 2007. Impact of duration and vowel inventory on formant values of oral vowels: An automated formant analysis from eight languages. *International Conference on Phonetics Sciences*, 1417–1420, Saarbrücken, Germany.

Wilbert Heeringa, Heike Schoormann, and Jörg Peters. 2015. Cross-linguistic vowel variation in saterland: Saterland Frisian, Low German, and High German. *Journal of the Acoustical Society of America*, 25–29.

Allard Jongman, Marios Fourakis, and Joan A. Sereno. 1989. The Acoustic Vowel Space of Modern Greek and German. *Language and Speech*, 32(3):221–248.

Andreas Kipp, Maria-Barbara WesenickM, and Florian Schiel. 1997. 2004): Maus goes iterative. *Proceedings of the Fifth European Conference on Speech Communication and Technology EUROSPEECH 1997*.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Chloé Larouche and François Steffann. 2018. Vowel space of French and Inuktitut: An exploratory study of the effect of vowel density on vowel dispersion. *Proceedings of the Workshop on the Structure and Constituency of Languages of the Americas*, vol. 21.

Wai-Sum Lee. 2012. A cross-dialect comparison of vowel dispersion and vowel variability. *2012 8th International Symposium on Chinese Spoken Language Processing*, 25–29.

Björn Lindblom. 1990. Explaining phonetic variation: A sketch of the h&h theory. *Speech Production and Speech Modelling*, 403–439. Springer Netherlands, Dordrecht.

Peter Livijn. 2000. Acoustic distribution of vowels in differently sized inventories - hot spots or adaptive dispersion? *PERILUS*, 93–96.

Christine Meunier, Cheryl Frenck-Mestre, Taissia Lelekov-Boissard, and Martine Le Besnerais. 2003. Production and perception of vowels: does the density of the system play a role? *HAL archives ouvertes*, 723–726. Université Autonome de Barcelone.

Daniel Recasens and Aina Espinosa. 2009. Dispersion and variability in catalan five and six peripheral vowel systems. *Speech Communication*, 51:240–258.

Elizabeth Salesky, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black, and Jason Eisner. 2020. A corpus for large-scale phonetic typology. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4526–4546, Online. Association for Computational Linguistics.

Florian Schiel. 2004. 2004): MAUS goes iterative. *Proceedings of the LREC 2004*, 1015–1018.

Hadley Wickham. 2017. tidyverse: Easily install and load the Tidyverse. *R package version 1.2.1*.

Raphael Winkelmann, Klaus Jaensch, Steve Cassidy, and Jonathan Harrington. 2021. emuR: Main Package of the EMU Speech Database Management System. *R package version 2.3.0*.

MERCI DE VOTRE ATTENTION!