Investigating phonological theories with crowd-sourced data: The Inventory Size Hypothesis in the light of Lingua Libre

Mathilde Hutin
Université Paris-Saclay
LISN-CNRS (UMR 9015)
Bât 507, 91405 Orsay, France
mathilde.hutin@lisn.fr

Marc Allassonnière-Tang
Muséum national d'Histoire naturelle
Laboratoire Eco-Anthropologie (UMR 7206)
17, place du Trocadéro, 75016 Paris, France
marc.allassonniere-tang@mnhn.fr

Abstract

Data-driven research in phonetics and phonology relies massively on oral resources, and access thereto. We propose to explore a question in comparative linguistics using an open-source crowd-sourced corpus, Lingua Libre, Wikimedia's participatory linguistic library, to show that such corpora may offer a solution to typologists wishing to explore numerous languages at once. For the present proof of concept, we compare the realizations of Italian and Spanish vowels (sample size = 5000) to investigate whether vowel production is influenced by the size of the phonemic inventory (the Inventory Size Hypothesis), by the exact shape of the inventory (the Vowel Quality Hypothesis) or by none of the above. Results show that the size of the inventory does not seem to influence vowel production, thus supporting previous research, but also that the shape of the inventory may well be a factor determining the extent of variation in vowel production. Most of all, these results show that Lingua Libre has the potential to provide valuable data for linguistic inquiry.

1 Introduction

One of the main challenges in data-driven research on the phonetics-phonology interface is the access to reliable, exploitable oral resources in sufficient amounts. While linguists working on other linguistic levels such as semantics or syntax can use written data as a proxy for language production, phoneticians and phonologists are limited to oral data, thus relying on audio recordings for vocal languages or video recordings for signed languages. Accessing massive amounts of such data is difficult enough, especially for studies in language comparison, that require such amounts in not one, but at the very least two languages.

To overcome this challenge, researchers developed two strategies. On the one hand, they can collect their own corpora, e.g., the CMU Wilderness Corpus (Black, 2019) or its emanation, the VoxClamantis corpus (Salesky et al., 2020), or other types of language-specific laboratory recordings such as the TIMIT database for English (Garofolo et al., 1993) or NCCFr for French (Torreira et al., 2010). On the other hand, they can gather audio recordings from other sources such as TV or radio shows, as was done for instance in the framework of the international project OSEO Quaero (www.quaero.org/), or from audio books, as exemplified by the LibriSpeech corpus for English (Panayotov et al., 2015, www.openslr.org/12). Both options have the disadvantage of being overly costly, both in money and human resources, and sometimes not freely accessible to the community. A third path has been recently explored: crowd-sourced data, recorded by volunteers and therefore much less costly in time and money and generally open-source. The project Common Voice (Ardila et al., 2020, https://commonvoice.mozilla.org) for instance was launched in 2017 by Mozilla for the intended purpose of creating a free database for the development of speech recognition software. In March 2022, it contains ∼18,000 hours of speech, 14,000 of which have been validated by other speakers, in 87 languages.

In the present paper, we explore a similar project: Lingua Libre, a participatory linguistic media library developed by Wikimedia France (https://lingualibre.org). It was launched in 2015, and, in March 2022, it counts ∼700,000 recordings in 148 languages across 775 speakers. This database is interesting to explore because it differs from Common Voice in the fact that its aim is not primarily the development of new technologies, or even linguistic inquiry in general, but patrimonial conservation of languages. Lingua Libre was used only once for academic purposes, i.e., to automatically estimate the transparency of orthographies in 17 languages (Marjou, 2021). With this study,

19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 23 - 28
July 14, 2022 ©2022 Association for Computational Linguistics

we aim to show that such data can be easily processed and useful to answer phonological questions in linguistic typology. In this proof of concept, we explore the realization of vowels by comparing two Romance languages: Italian and Spanish.

The outline of the paper is as follows. In Section 2, we describe our research question to justify our choice of languages. In Section 3, we present our corpus and methodology. In Section 4, we provide an analysis of the vowels in Italian and Spanish. Section 5 concludes and discusses the results.

2 The Inventory Size Hypothesis *vs* the Vowel Quality Hypothesis

In this paper, we offer to use Lingua Libre to tackle the question of vowel production with regards to vowel inventory. Our research question stems from various theories regarding the shape of vowel inventories in the world's languages. Our study however focuses on synchronic phonetic variation with regards to phonological systems (on the phylogeny of vowel systems in the languages of the world, see Zhang and Gong (2022) and references therein).

The original Vowel Dispersion Theory (Liljencrants and Lindblom, 1972; Lindblom, 1986) and a few years later the Adaptive Dispersion Theory (Lindblom, 1990), stem from the H&H (”Hypo- and Hyperspeech”) model of communication, that assumes that speakers tend toward minimal and sufficient perceptual contrast, i.e., operate a trade-off between articulatory economy (hypospeech) and perceptual distinctiveness (hyperspeech). In the original works, these theories are the foundation for phylogenetic research on the distribution of vocalic categories in the languages of the world, for instance to explain why three-vowel systems usually display /a, i, u/ and not, say, /a, y, u/. Phoneticians however have particularly focused on one hypothesis that emerges from this model: The more vocalic categories the language has in its phonemic inventory, the less phonetic variation the corresponding vowel realizations will display. This is the hypothesis we ourselves focus on in the present paper, to which we will refer as the Inventory Size Hypothesis, henceforth ISH.

This hypothesis has been tested in a number of studies, with contradictory results. Jongman et al. (1989) on American English, Greek and German, Al-Tamimi and Ferragne (2005) on French and two dialects of Arabic and Larouche and Steffann (2018) on Quebec French and Inuktitut support the ISH while Bradlow (1995) on English and Spanish, Meunier et al. (2003) on English, Spanish and French, Recasens and Espinosa (2009) on 5 dialects of Catalan, Lee (2012) on 5 dialects of Chinese and Heeringa et al. (2015) on 3 German languages, do not provide evidence in favor of the ISH, which can be due, for the last three at least, to the genetic and geographical closeness of the languages and possible bilingualism of the speakers. Studies on larger sets of languages however tend to invalidate the hypothesis: Engstrand and Krull (1991) found inconclusive results on 7 languages across 6 language families; Livijn (2000) on 28 languages, Gendrot and Adda-Decker (2007) on 8 languages across 4 families, and Salesky et al. (2020) on 38 languages across 11 families, found no evidence for an effect of inventory size on the global acoustic space.

Building on these negative results, we suggest that it may not so much be the number of categories but their actual quality that influences the vowel's realizations. For instance, between two imaginary languages A and B displaying /a, e, i, o, u/ *vs* /a, e, i, y, o, u/ respectively, it is also possible that not all the categories in language B will display less variation than those in language A: Only [i] and possibly [u], which compete with /y/ in B but not in A, would show less variation in B than in A. We propose to refer to this restatement of the original hypothesis, as the Vowel Quality Hypothesis, henceforth VQH.

In this paper, we aim to test this alternative: Either the ISH is valid, and all the vowels of the system will be affected by the size of the inventory, or the VQH is more accurate, and only some vowels or some acoustic parameters will be affected depending on the other vowels comprised in the system. The third possible outcome is that neither the ISH nor the VQH is accurate.

To test our hypothesis, we focus on the F1 and F2 values of the vowels in two Romance languages: Spanish and Italian. Spanish has a limited vowel inventory, with only 5 categories /a, e, i, o, u/ while Italian has 7: /a, ɛ, e, i, o, ɔ, u/. Their inventories differ only in the number of degrees of aperture (Spanish has open, mid and closed vowels while Italian has open, mid-open, mid-closed and closed vowels), which manifest as variation on the first frequency, F1. If the ISH is valid, we expect vowel productions from each language to differ in both F1 and F2, while if the VQH is valid, we expect Spanish and Italian vowels to differ only in F1.

3 Materials and Methodology

As a crowd-sourcing tool, Lingua Libre allows any speaker to log in, fill in a profile with basic metadata for themselves of for other speakers, and record themselves or their guests reading lists of words in their language. The device detects pauses, which allows for the recording to end when the word has been read and the next recording to start automatically after, therefore effortlessly generating relatively short audio files for each word. Each audio file is supposed to be titled on the same template of ‘Language - Speaker - Item’. For example, for the recording ‘spa.-Marreromarco-solucionar.wav’, the language is Spanish (‘spa’), the speaker ID is ‘Marreromarco’, and the recorded item is ‘solucionar’, ‘solve’. All audio files are under a Creative Commons licence, i.e., open-source.

First, the recordings are scrapped from the Lingua Libre database. In the present study, we extract a subsample of 500 items for /a, e, i, o, u/ in each language, to counter the fact that both languages have different amounts of data points and to also control for number of speakers (5) in each language. In total, we have 500 occurrences for each of the 5 vowels in both Italian and Spanish, which results in 5000 tokens. To avoid a potential sample bias, the sampling of tokens is conducted 10 times. We also took care to limit our investigation to the European variety of Spanish, to avoid any mismatch with the more limited geographical expansion of Italian.

Second, the recordings are segmented and aligned using WebMAUS (Kisler et al., 2017), the online open-access version of the MAUS software (Schiel, 2004). MAUS creates a pronunciation hypothesis graph based on the orthographic transcript of the recording (extracted from the name of the audio file) using a grapheme-to-phoneme converter. During this process, the orthographic transcription is converted to the Speech Assessment Methods Phonetic Alphabet (SAMPA). The signal is then aligned with the hypothesis graph and the alignment with the highest probability is chosen. Experiments have shown that the MAUS-based alignment is 95% accurate compared to human-based alignments (Kipp et al., 1997).

Third, the selected vowels are extracted from the recordings and analyzed in terms of formants. For each recording of each vowel, the mean F1 and F2 of the entire sound are calculated. The mean formants are considered to attenuate the effect of co-articulation with the left and right contexts.

Vowel	a	i	o
ID	9309	4238	48269
iso	ita	ita	spa
F1	664	315	628
F2	1451	2494	1153
Speaker	LangPao	LangPao	Rodelar
Item	rosa	chimica	todo

Table 1: Example of the data extracted and compiled from Lingua Libre. Each column represents one data point.

Table 1 shows an example of the extracted and compiled data used in this study. Each occurrence of vowel is given a unique identifier to allow tracking it within a word that has several vowels. The language iso code is provided along with the values of F1 and F2. Finally, the recorded word and its contributor are also noted. For the whole process, the following R packages are used: `emuR` (Winkelmann et al., 2021), `PraatR` (Albin, 2014), and `tidyverse` (Wickham, 2017).

4 Results: Shape of the inventory, more than size, influences vowel production

We focus on the F1 and F2 values for the 5 vowels that Spanish and Italian have in common, /a, e, i, o, u/. Our hypothesis is that, if the ISH is valid, we will find variation in both F1 and F2 for all vowels, while if the VQH is valid, we will find variation only in F1, especially in /a/, /e/ and /o/, which are in direct competition with /ɛ/ and /ɔ/.

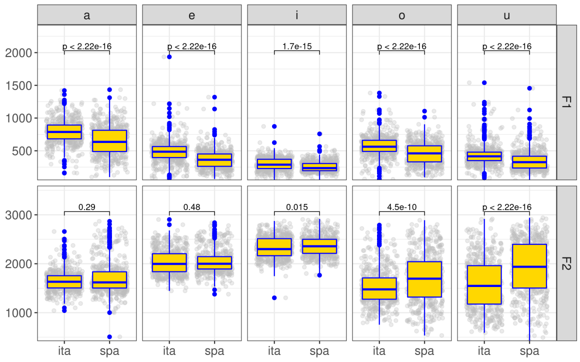

Figure 1: Distribution of formants for each of the 500 [a], [e], [i], [o], and [u] across the Italian and Spanish data extracted from Lingua Libre. The significance labels indicate the output of a wilcoxon test with bonferroni correction.

As general information, Figure 1 provides the mean values for F1 (top tier) and F2 (bottom tier) in Italian (left brackets) and Spanish (right brackets) for all 5 vowels of interest. It shows that F1 is significantly lower in Spanish for all 5 vowels, while F2 is statistically higher only for back vowels.

To test our hypotheses, however, we are less interested in F1 and F2 values in general than in their variation. Figure 2 shows the variation coefficient (standard deviation divided by the mean) of F1 (top tier) and F2 (bottom tier) for each replication of each vowel category in Italian (left brackets) and Spanish (right brackets). Each point represents the variation coefficient of a formant and a vowel for a replication. These results show that there is significantly less variation in F1 in Italian /a/, /e/, /o/ and /u/ than in Spanish, thus supporting the VQH. The difference between F2 variation coefficients is also significant but inverted for /e/, /i/, and /u/ where we observe more variation for Italian than for Spanish, thus invalidating the ISH.

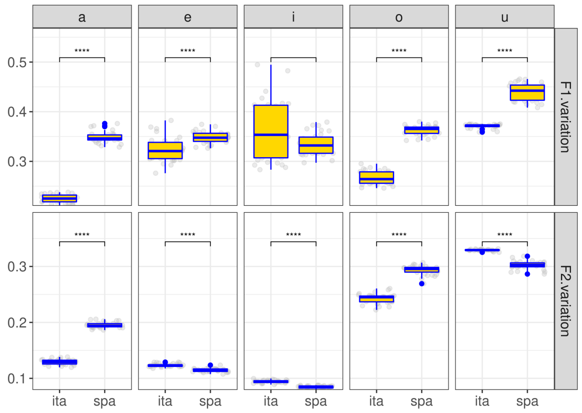

Figure 2: The distribution of the variation coefficient for each of the 500 [a], [e], [i], [o], and [u] across the Italian and Spanish data extracted from Lingua Libre in each of the replications. The significance labels indicate the output of a wilcoxon test with bonferroni correction.

These results are also supported by the linear mixed models we conducted (in both Bayesian and non-Bayesian versions) based on the 500 data points from each of the 10 replications. First, Table 2 shows that the estimate for the variation of Spanish for F1 is five times larger than the one for F2. Furthermore, we also observe that the variation is generally larger for most of the vowels in F1 (except for /a/), while the variation varies for F2, in which the estimates are negative for /e/ and /i/. The same observation is found when comparing the overall areas covered by the polygons formed by the contours of F1 and F2. We conduct a 2D kernel density estimation (Venables and Ripley, 2002) to extract the contours of the area covered by the occurrences of each vowel in the two-dimensional space from F1 and F2. While there is generally more variation in Spanish than in Italian, this varies across vowels, as /e/ and /i/ tend to have a smaller formant space in general.

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.

5 Conclusion and discussion

We used crowd-sourced data to test two competing hypotheses in language typology: The production of vowels is influenced either by the size of the inventory, or by its shape. Our proof-of-concept on Italian and Spanish shows that the size of the inventory does not influence the realization of vowels, but the exact quality of the vowels at hand does.

Our study also points to several caveats. First, all audio files were not properly labeled and were thus unusable. Moreover, from a human point of view, it should be noted that crowd-sourced data heavily rely on the participants' good will and that researchers have no choice but to trust the provided metadata. One possible solution to that last problem would be for Lingua Libre to propose a verification tool, as does Common Voice, to improve the reliability of the data and metadata. However, crowd-sourced data proved to be a promising tool for linguistic inquiry, especially to investigate language universals, and could thus be tested on more substantial sets of languages.

Acknowledgments

This research was partially supported by Institut DATAIA and the MSH Paris-Saclay in the framework of the Excellency Award for the project OTELO - OnTologies pour l'Enrichissement de l'analyse Linguistique de l'Oral (PI Ioana Vasilescu and Fabian Suchanek), and by the French National Research Agency in the framework of the grant EVOGRAM: The role of linguistic and nonlinguistic factors in the evolution of nominal classification systems, ANR-20-CE27-0021 (PI Marc Allassonnière-Tang). The authors would also like to thank the Wikimedia community for their interest in the project, and in particular Lucas Lévêque for his help on the Lingua Libre tool.

References

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

Aaron Albin. 2014. Praatr: An architecture for controlling the phonetics software "praat" with the r programming language. *Journal of the Acoustical Society of America*, 135(4):2198.

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. In *Proceedings of LREC*.

Alan W Black. 2019. Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.

Ann R. Bradlow. 1995. A comparative acoustic study of english and spanish vowels. *The 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.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue. 1993. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*.

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. In *International Conference on Phonetics Sciences*, pages 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. *The Journal of the Acoustical Society of America*, pages 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. In *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. In *Proceedings of the Workshop on the Structure and Constituency of Languages of the Americas*, volume 21.

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

Johan Liljencrants and Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48(4):839–862.

Björn Lindblom. 1986. Phonetic universals in vowel systems. *Experimental Phonology*, pages 13–44.

Björn Lindblom. 1990. Explaining phonetic variation: A sketch of the h&h theory. In William J. Hardcastle and Alain Marchal, editors, *Speech Production and Speech Modelling*, pages 403–439. Springer Netherlands, Dordrecht.

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

Xavier Marjou. 2021. Oteann: Estimating the transparency of orthographies with an artificial neural network. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 1–9. Association for Computational Linguistics.

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? In *hal archives ouvertes*, pages 723–726. Université Autonome de Barcelone.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4526–4546, Online. Association for Computational Linguistics.

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

Francisco Torreira, Martine Adda-Decker, and Mirjam Ernestus. 2010. The nijmegen corpus of casual french. *Speech Communication*, 52:201–212.

W. N. Venables and Brian D. Ripley. 2002. *Modern applied statistics with S*, 4th ed edition. Statistics and computing. Springer, New York. OCLC: ocm49312402.

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.

Menghan Zhang and Tao Gong. 2022. Structural variability shows power-law based organization of vowel systems. *Frontiers in Psychology*, 13.