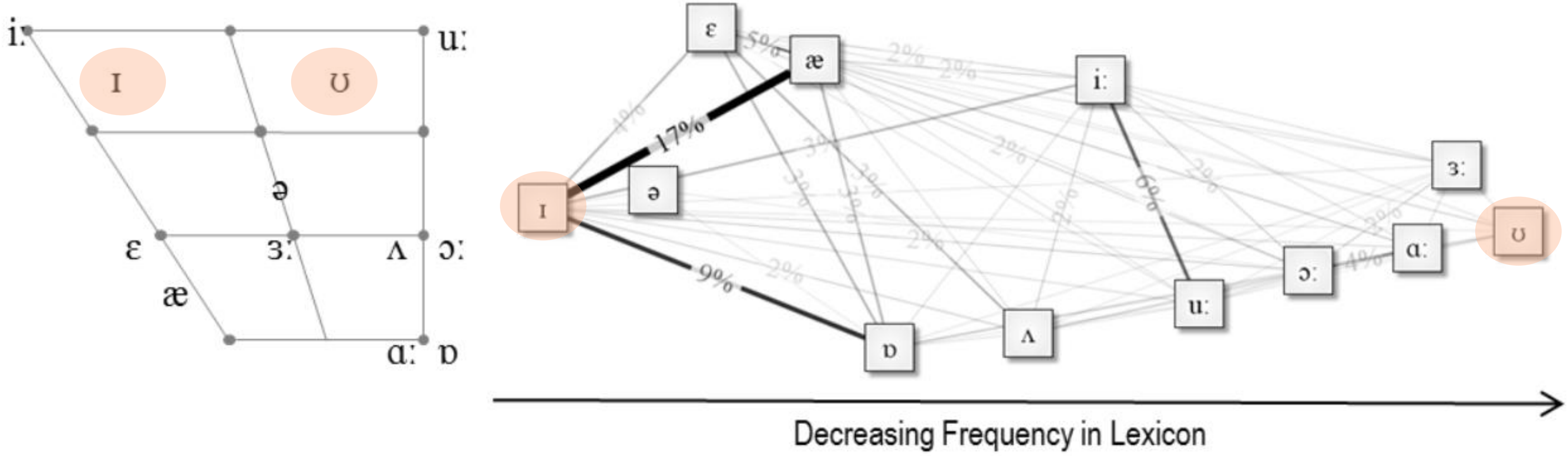
The nature of phonological contrasts as a function of their position within syllables and words

Yoon Mi Oh, Christophe Coupé & François Pellegrino
Ajou University, The University of Hong Kong &
Laboratoire Dynamique du Langage (CNRS; University of Lyon)

ICL21, September 10, 2024

Background



Illustrations of the English (Received Pronunciation) vowel system

Standard IPA chart (left) and Functional network-based representation (right)
Vowels are ranked from left to right by decreasing usage frequency.

> **“The function of a phonemic system is to keep the utterances of a language apart” but “some contrasts between the phonemes in a system apparently do more of this job than others.”**
>
> Hockett, C. F. (1966). The quantification of functional load: A linguistic problem, *Report Number RM-5168-PR,* Rand Corp., Santa Monica.

General framework

Phonological systems

- are characterized by the phenomenon of self-organization

- are organized by their systemic usage for efficient communication in the mental lexicon

- are quantitatively described using functional load (FL) in the search for cross-linguistic regularities

Oh, Y. M., Coupé, C., Marsico, E., & Pellegrino, F. (2015). Bridging phonological system and lexicon: insights from a corpus study of functional load. *Journal of Phonetics*, 53.

General framework

Functional load (FL)：

estimates the relative importance of a phoneme contrast (x, y) by quantifying the perturbation induced by merging a pair of phonemes x and y in terms of homophony and lexical informativeness of the lexicon.

- **is considered as a key predictor of sound change in historical linguistics**

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition, 128*(2), 179-186.

- **explains various linguistic phenomena in first and second language acquisition and language perception**

Lin, I. (2019). *Functional load, perception, and the learning of phonological alternations*. University of California, Los Angeles.

- **is used as a tool for cross-linguistic description of phonological systems and mental lexicon in linguistic typology**

Oh, Y. M., Coupé, C., Marsico, E., & Pellegrino, F. (2015). Bridging phonological system and lexicon: insights from a corpusstudy of functional load. *Journal of Phonetics*, 53.

Research questions

- A recent study has shown that onsets play a more important role than codas in keeping words distinct by comparing the number of minimal pairs in these two positions (Sun & Poeppel, 2023).

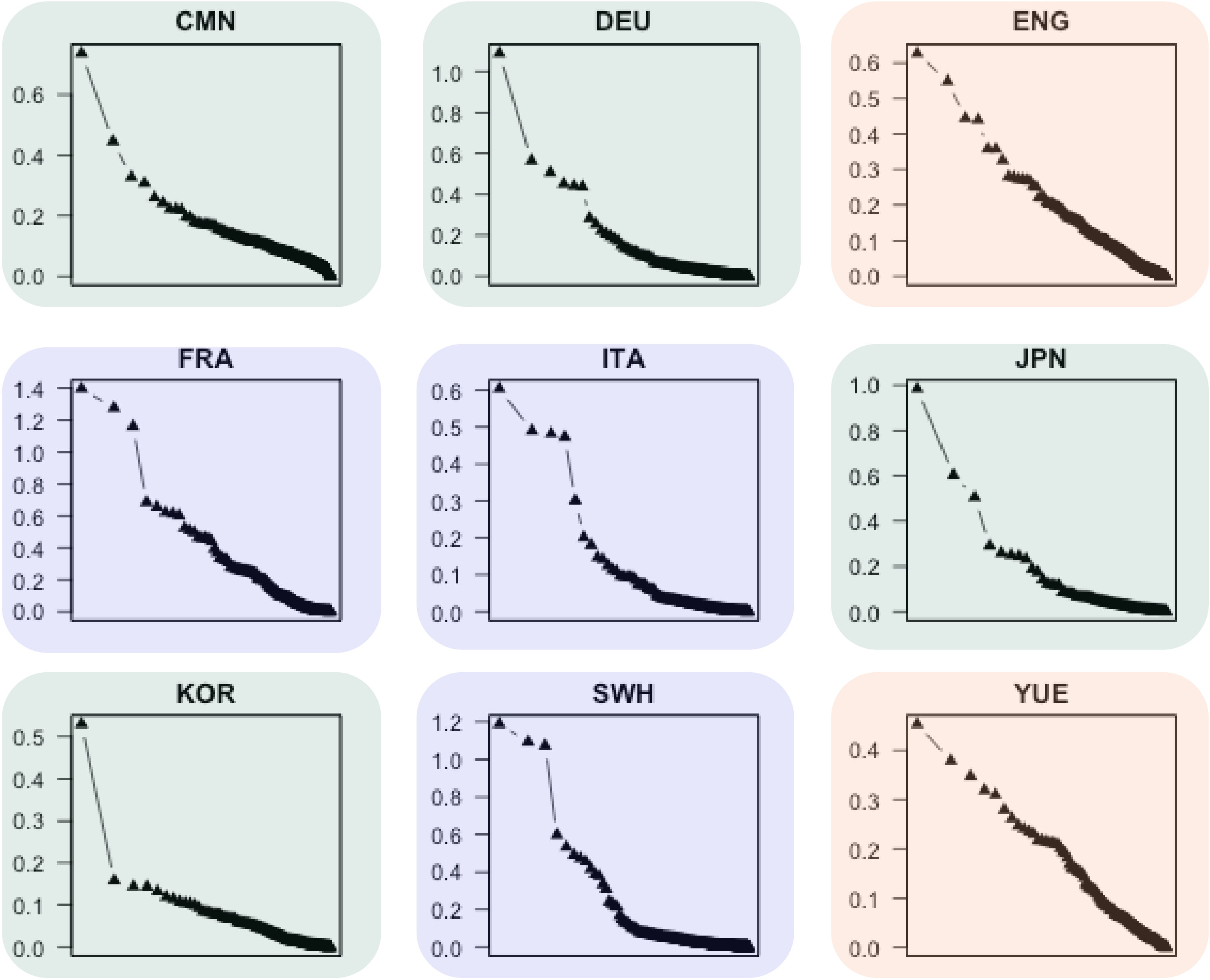
- How does such a functional asymmetry manifest itself phonologically? How does the position of a phonemic contrast within syllables and within words affect the phonological distances involved?

→ By looking at (i) the distribution of phonological distance and (ii) the relationship between phonological distance and functional load within syllables and within words

Previous works (I)

FL distribution of consonant pairs (y-axis) in nine languages

Pairs are listed in descending order of FL values using a semi-logarithmic scale on the x-axis.



- **Uneven distribution of FL: only few consonant contrasts play a major role in differentiating words.**

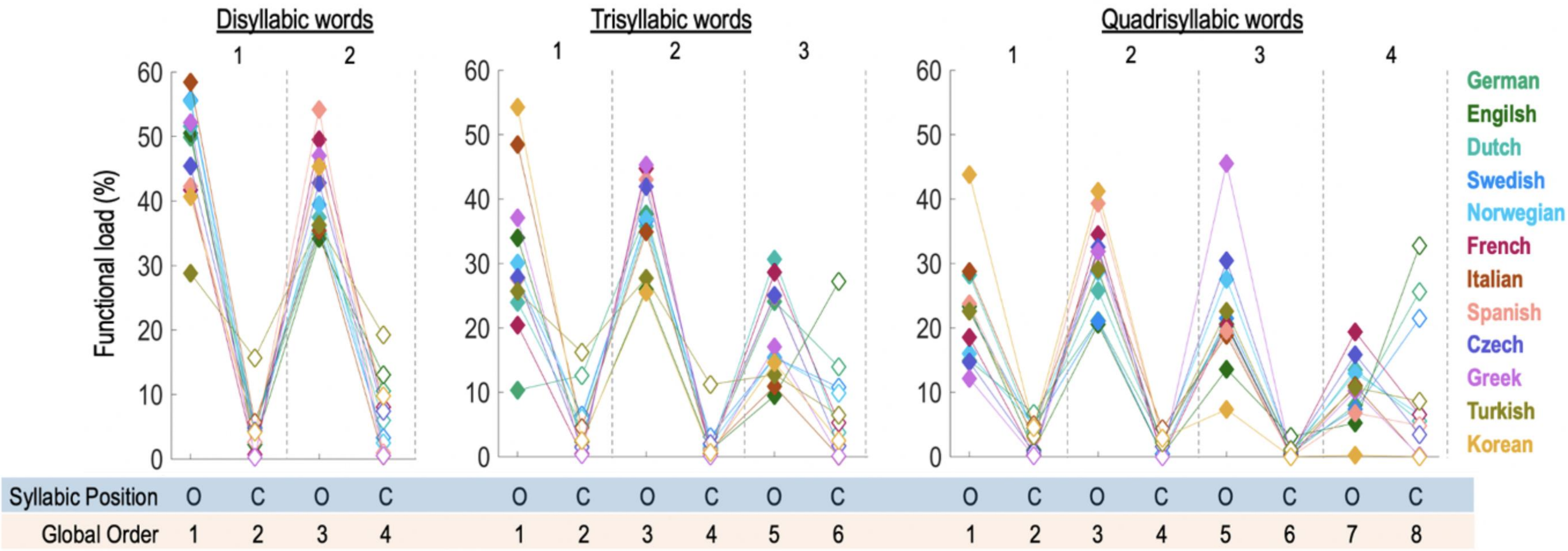
→ robustness and resilience to errors

- **Language-specific differences are visible.**

Oh, Y. M., Coupé, C., Marsico, E., & Pellegrino, F. (2015). Bridging phonological system and lexicon: insights from a corpus study of functional load. *Journal of Phonetics*, 53.

Previous works (II)

Variation of FL across the onset and coda positions of each syllable in multi-syllable words in 12 languages

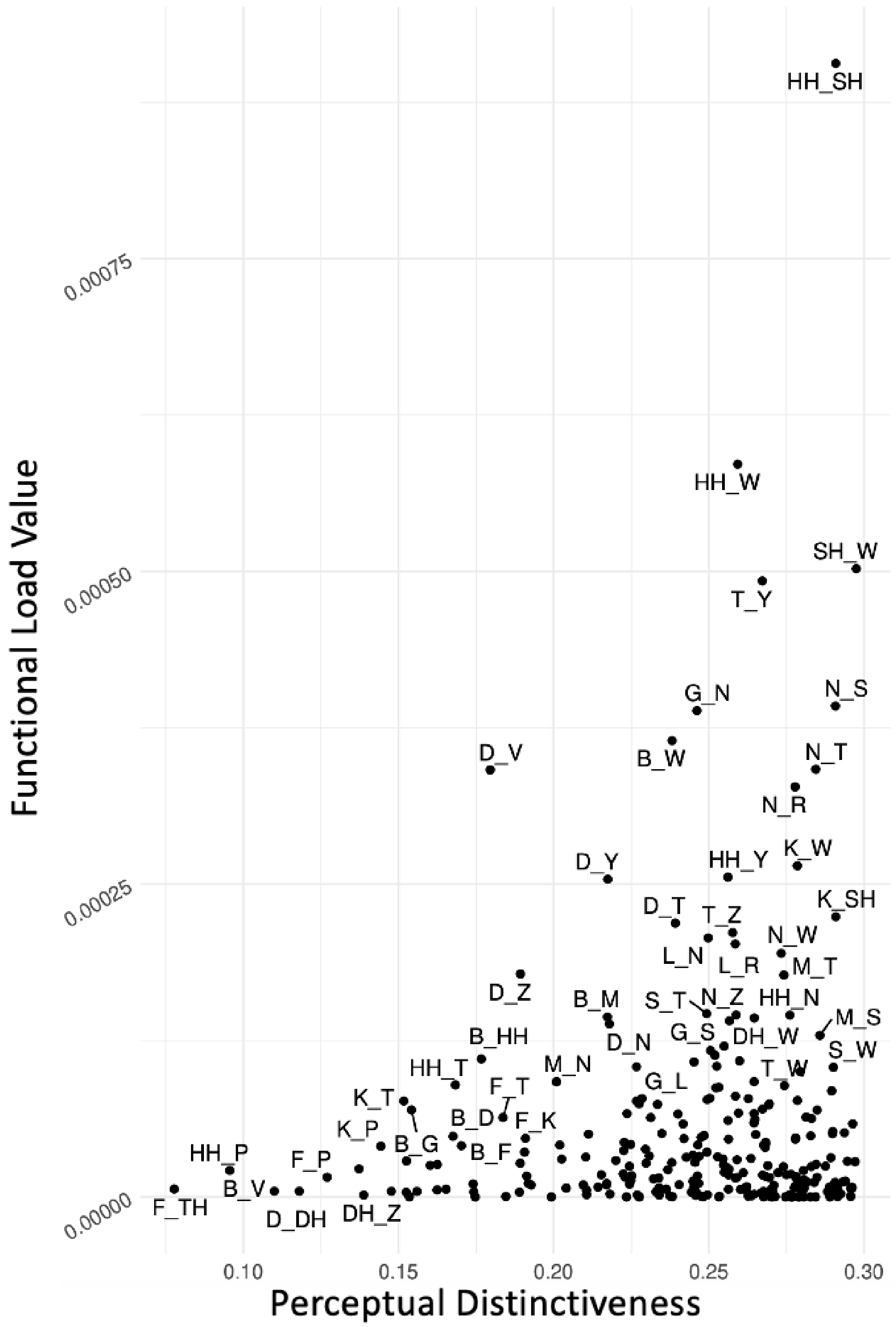


- A higher FL of onsets than of codas in twelve languages ➡ Functional asymmetry between onsets and codas for lexical informativeness. Such an asymmetry results from the modulation of lexical informativeness within the same syllable unit, not from an overall decrease of FL at the whole word level.
- Most of the sampled languages show a strong preference for suffixation with more morphological information at the end of words or syllables. ➡ Analytic and isolating languages (e.g., Māori and Vietnamese) should be addressed

Sun, Y., & Poeppel, D. (2023). Syllables and their beginnings have a special role in the mental lexicon. *Proceedings of the National Academy of Sciences, 120*(36), e2215710120.

Previous works (III)

FL for 276 English consonant pairs (y-axis) and perceptual distinctiveness (x-axis)



- **The FL for a consonant contrast increases significantly with its perceptual distinctiveness (β = 0.13, t = 3.45, p < 0.001)**

- **Phonological contrasts of higher perceptual distinctiveness do more work in keeping words distinct due to communicative pressures to minimize the likelihood of perceptual confusion.**

Zhang, Y., Li, Z., Wu, B., Xie, Y., Lin, B., & Zhang, J. (2021). Relationships between perceptual distinctiveness, articulatory complexity and functional load in speech communication. *Interspeech*, 1733-1737.

Data and language (FL_E)

Language	ISO 639-3 Code	Data
Basque	EUS	**E-Hitz** Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E., & Carreiras, M. (2006). E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods, 38*(4), 610-615.
French	FRA	**Lexique 3.80** New,B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE 3.80, *L'Année Psychologique, 101*, 447-462.
German	DEU	**WebCelex** Max Planck Institute for Psycholinguistics, *WebCelex*, retrieved on March 18, 2013 and on August 6, 2014 from http://celex.mpi.nl.
Italian	ITA	**PAISÀ Corpus** Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., & Pirrelli, V. (2014). The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, Association for Computational Linguistics, Gothenburg, Sweden, 36-43.
Korean	KOR	**Leipzig Corpora Collection** Universität Leipzig, Leipzig corpora collection (LCC) retrieved on 2013 from http://corpora.informatik.uni-leipzig.de.
Māori	MRI	**Māori Broadcast Corpus** Boyce, M. T. (2006). A corpus of modern spoken Māori. *Unpublished doctoral thesis in Applied Linguistics.* Victoria University of Wellington. **MAONZE Corpus** King, J., Maclagan, M., Harlow, R., Keegan, P., & Watson, C. (2011). The MAONZE project: Changing uses of an indigenous language database. *Corpus Linguistics and Linguistic Theory, 7*(1), 37-57.

Data preprocessing (FL_E)

Text corpus

- are mostly retrieved online from different sources
- the 30k most frequent word forms are considered except for Italian (15,788 word forms).
- Different strategies for preprocessing the text corpus

Automatic phonological transcription and syllabification

- For 3 languages (EUS, FRA, DEU), processed data already available

- For 3 languages (ITA, KOR, MRI) : phonologically transcribed according to language-specific transcription rules and automatically syllabified by a program written by the author

Data and language (FL_{MP})

Language	ISO 639-3 Code	Data
English	ENG	**WebCelex** Max Planck Institute for Psycholinguistics, *WebCelex*, from http://celex.mpi.nl.
French	FRA	**Lexique 3.81** New,B., Pallier, C., & Ferrand, L. (2005). La documentation officielle de Lexique 3.
German	DEU	**WebCelex** Max Planck Institute for Psycholinguistics, *WebCelex*, from http://celex.mpi.nl.
Italian	ITA	**PhonItalia1.10** Goslin, J., Galluzzi, C., & Romani, C. (2014). PhonItalia: a phonological lexicon for Italian. *Behavior Research Methods, 46*, 872-886.
Korean	KOR	**K-SPAN** Holliday, J. J., Turnbull, R., & Eychenne, J. (2017). K-SPAN: A lexical database of Korean surface phonetic forms and phonological neighborhood density statistics. *Behavior Research Methods, 49,* 1939-1950.
Spanish	SPA	**BuscaPalabras** Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods, 37,* 665-671.

- **Number of word forms varying between languages from 26k(SPA) to 50k(DEU)**
- **Data preprocessing: phonological transcription and syllabification provided by the dataset**

Methodology (I)

● Minimal pair-based definition of FL

FL*MP****(x, y)***: number of different word forms distinguished (only) by the contrast between the phonemes *x* and *y*

Ingram, D. (1989). *First language acquisition: Method, description and explana*tion. Cambridge University Press.

cf. Phonological neighbor: word forms differed by substitution, addition or deletion of a phoneme

*FL**position*: relative proportion (%) of minimal pairs for each syllable position (onset and coda) among all identified minimal pairs for each contrast

$$FL_{position} = \frac{MP_{position}}{MP_{onset} + MP_{coda}} \times 100\%$$

Sun, Y., & Poeppel, D. (2023). Syllables and their beginnings have a special role in the mental lexicon. *Proceedings of the National Academy of Sciences, 120*(36), e2215710120.

● Entropy-based definition of FL

FL*E****(x, y)***: relative difference in entropy between the observed state *L* and a fictive state L^*_{xy} in which the contrast between the phonemes *x* and *y* is neutralized

Hockett, C. F. (1966). The quantification of functional load: A linguistic problem, *Report Number RM-5168-PR,* Rand Corp., Santa Monica.

$$FL_E(x,y) = \frac{H(L) - H({L^*}_{xy})}{H(L)}$$

Language *L* considered as a source of sequences of independent words w_i taken from a set N_L

Methodology (II)

- Entropy-based calculation of FL

Observed Lexicon

Form	Frequency
pal	300
pil	200
bal	150
bil	150
pul	100
bul	100
TOTAL	1000

Inventory: /a i u p b l/

$N_L = 6$ $H(L) = 2.47$

Stress taken into account for EUS, DEU, ITA.

Contrast /a-i/

Form	Frequency
p*l	300
p*l	200
b*l	150
b*l	150
pul	100
bul	100
TOTAL	1000

Fictive Lexicon

Form	Frequency
p*l	500
b*l	300
pul	100
bul	100
TOTAL	1000

$H(L^*_{ai}) = 1.69$

FL(a-i) = (2.47-1.69)/2.47 = 31.6 %

FL(a-u) = 23.1 %

FL(i-u) = 21.0 %

Phoneme /a/

$$FL(x) = \frac{1}{2}\sum_{y} FL(x,y)$$

FL(a) = ½ (FL(a-i)+FL(a-u)) = ½ (31.6+23.1) = 27.35 %

FL_V=61%

Methodology (III)

- FL_E calculated for each syllable position within words

(1) The unique syllables of monosyllabic words (Mono): S_{Mono}

(2) The first syllables of disyllabic or longer words (First): $S_{\text{First}}\_S_2$, $S_{\text{First}}\_S_2\_S_3$

(3) The middle syllables of trisyllabic or longer words (Mid): $S_1\_S_{\text{Mid}}\_S_3$, $S_1\_S_{\text{Mid}}\_S_{\text{Mid}}\_S_4$

(4) The last syllables of disyllabic or longer words (Last): $S_1\_\ S_{\text{Last}}$, $S_1\_S_2\_S_{\text{Last}}$

- FL_{MP} for consonant pairs calculated for each onset and coda position within syllables and within words

Methodology (IV)

- **Phonological distance is calculated by replicating the method described in the panphon.distance module of the PanPhon package based on 22 subsegmental articulatory features coded as +, - or 0 in the PanPhon database of more than 6,000 IPA segments.**

- **Phonological distance is defined as an edit distance where each feature-edit has a cost of $^{1}/_{22}$.**

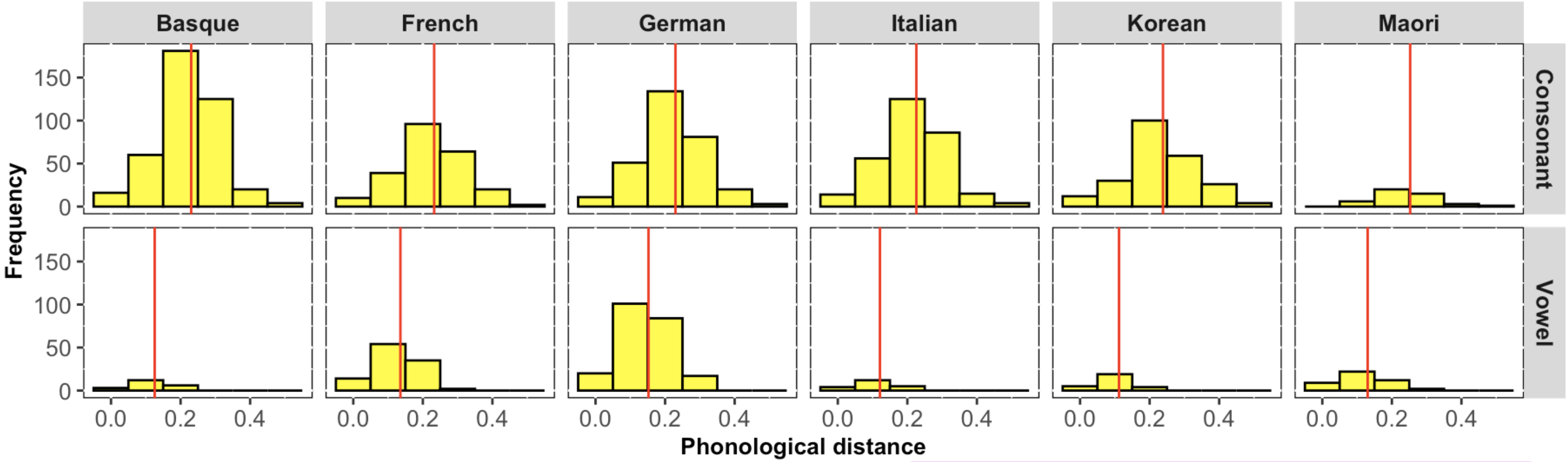
	syl	son	cons	cont	delrel	lat	nas	strid	voi	sg	cg	ant	cor	distr	lab	hi	lo	back	round	velaric	tense	long
/p/	-	-	+	-	-	-	-	0	-	-	-	+	-	0	+	-	-	-	-	-	0	-
/p^h/	-	-	+	-	-	-	-	0	-	+	-	+	-	0	+	-	-	-	-	-	0	-
/p^j/	-	-	+	-	-	-	-	0	-	-	-	+	-	0	+	+	-	-	-	-	0	-
/p^{hj}/	-	-	+	-	-	-	-	0	-	+	-	+	-	0	+	+	-	-	-	-	0	-

Phonological distance between /p^h/ and /p^j/ = $^{1}/_{22}$ + $^{1}/_{22}$ = $^{1}/_{11}$

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C. & Levin, L. (2016). PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,* 3475-3484.

Results (I)

Distribution of phonological distance between pairs of consonants and pairs of vowels



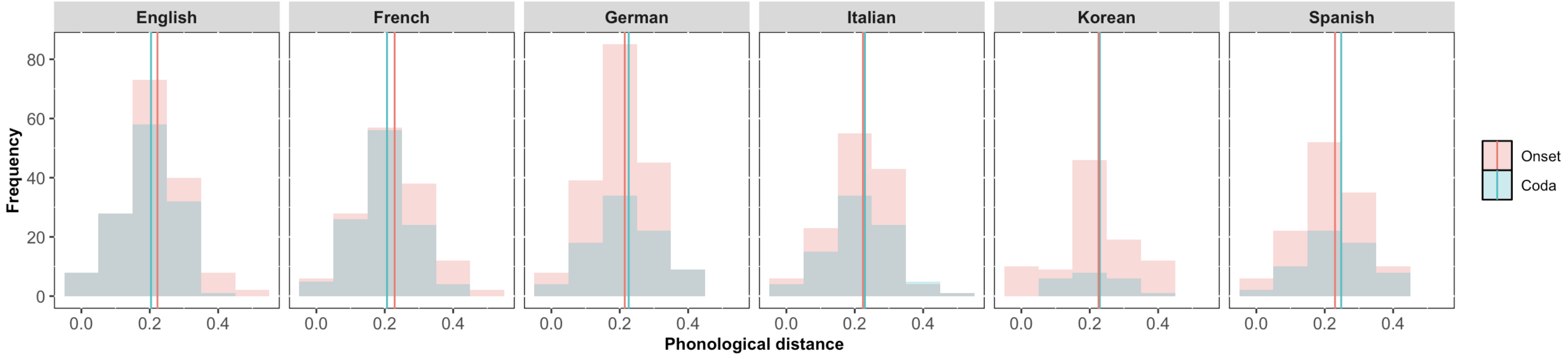
The average value of 0.232 corresponds to having 5.1 different articulatory features out of 22.

- On average, a larger average phonological distance for consonant pairs (min = 0; max = 0.52; mean = 0.232; SD = 0.093) than for vowel pairs (min = 0; max = 0.32; mean = 0.141; SD = 0.064) in six languages
- A cross-linguistic tendency for "mid" distances

The average value of 0.141 corresponds to having 3.1 different articulatory features out of 22.

Results (II)

Distribution of phonological distance between onset and coda pairs

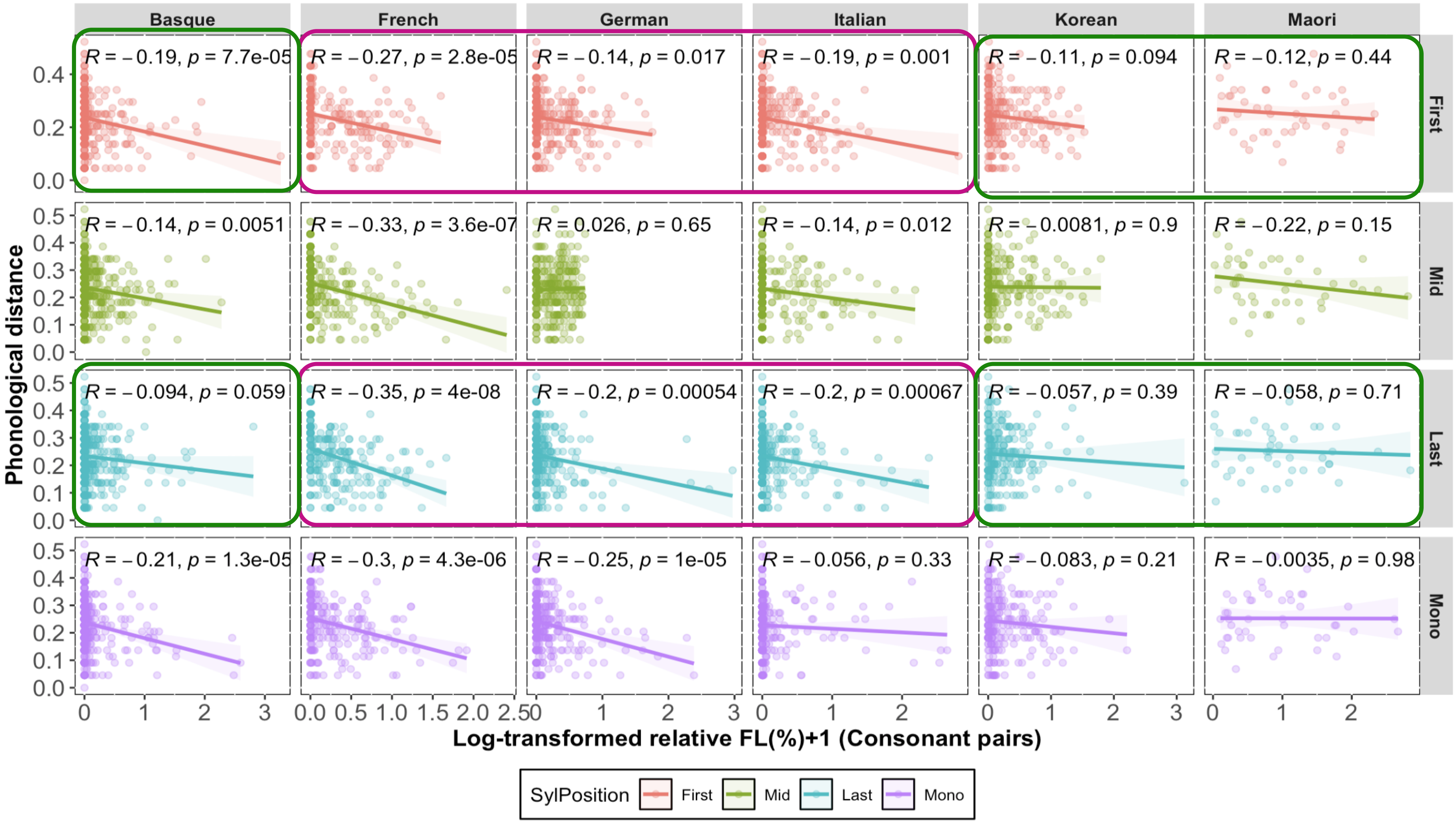


- No cross-linguistic trend in differences between average phonological distances for onset (mean = 0.224; SD = 0.092) and coda pairs (mean = 0.219; SD = 0.091) across six languages

The average value of 0.224 corresponds to having 4.93 different articulatory features out of 22 and 0.219 corresponds to 4.82 different articulatory features.

Results (III)

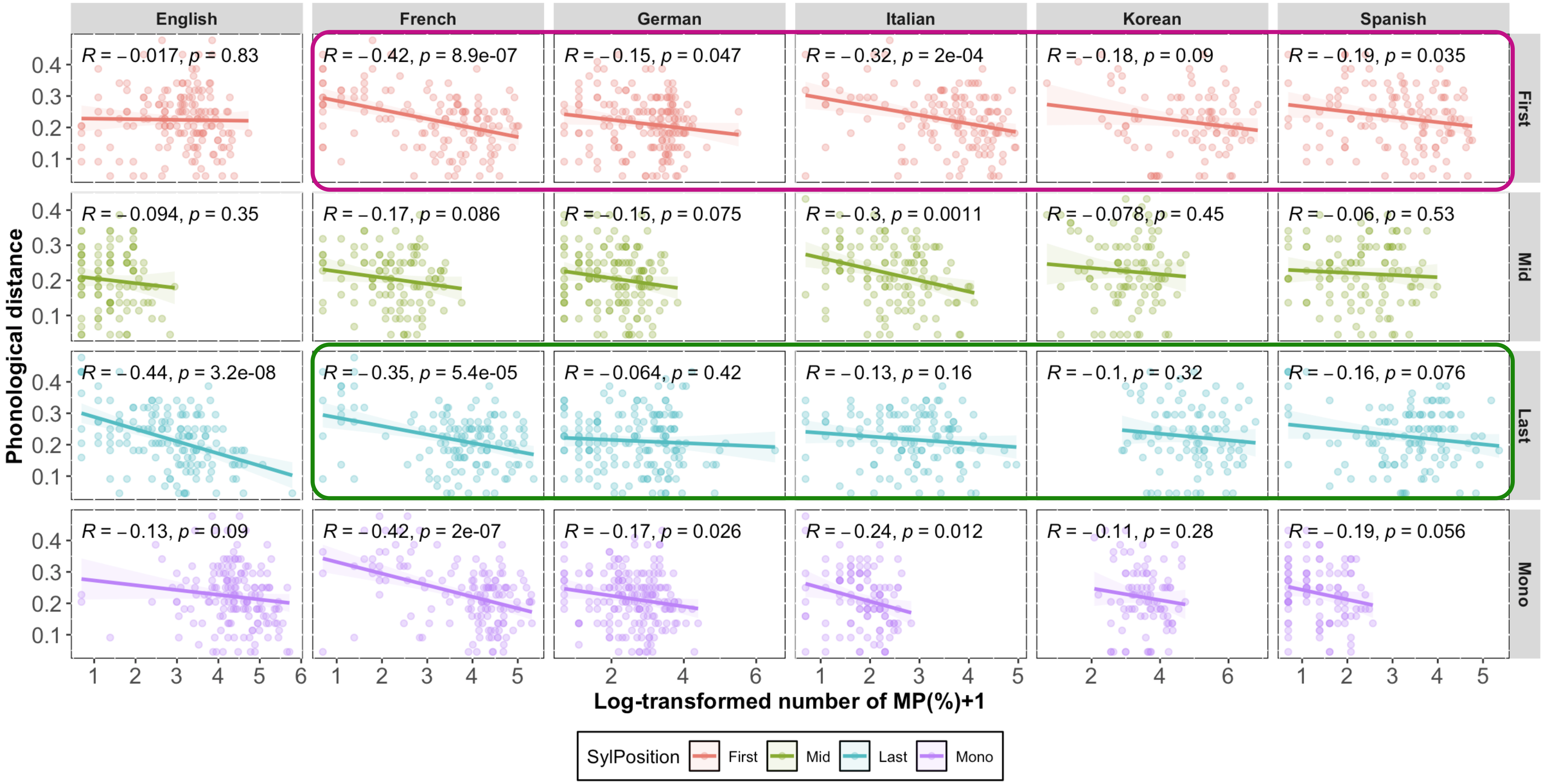
Phonological distance and relative FL_E of consonant contrasts within words in six languages



- In contrast to agglutinative (Basque and Korean) and analytic (Māori) languages, fusional languages (French, German, and Italian) exhibit a stronger negative tendency in the last syllables than in the first ➔ cross-linguistic tendency for languages with a strong suffixation

Results (IV)

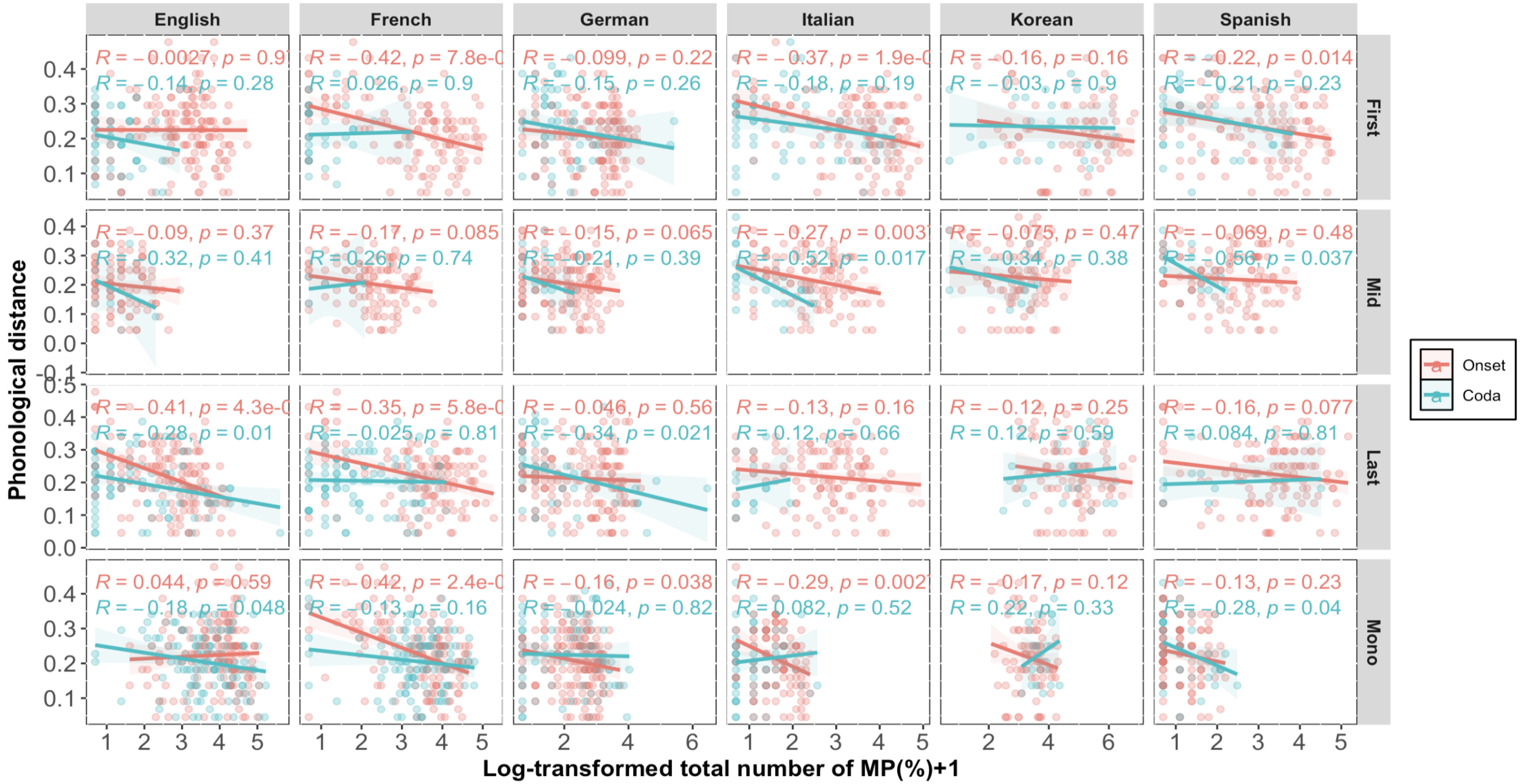
Phonological distance and FL_{MP} of consonant contrasts within words in six languages



- However, when word frequencies are not considered (as measured by FL_{MP}), there is a weaker negative tendency or correlation in the last syllables than in the first syllables in most languages except English ➡ cross-linguistic tendency to improve intelligibility in the last syllables by minimizing the phonological distance

Results (V)

Phonological distance and $\mathrm{FL}_{\mathrm{MP}}$ of consonant contrasts within syllables and within words in six languages



- Similar to FL_{E}, the number of minimal pairs is weakly negatively correlated with phonological distance in six languages (***β*** = -0.05, ***t*** = -3.0835, ***p*** = 0.002059) . However, no cross-linguistic tendency is observed at any position within syllables and within words.

Conclusions

- Quite large phonological distances for both consonant and vowel pairs in all six languages

- Cross-linguistic preference for “mid” differences ➔ Symmetric distribution of phonological distances between phonemic contrasts across all six languages with no visible position-specific tendency

- Negative correlation between relative FL_E and its phonological distance ➔ No cross-linguistic preference for phonemic pairs of higher functional load at a larger phonological distance

- Effect of position within words: weaker negative correlation in the last syllables than in the first syllables except English ➔ cross-linguistic tendency to improve intelligibility in the last syllables by minimizing the phonological distance

- Effect of position within syllables ➔ no cross-linguistic tendency observed

Perspectives

- Consider more typologically diverse languages

➔ including agglutinative, analytic and isolating languages

- Uniformize the data size (i.e. number of word forms) and apply the two calculation methods (FL_E and FL_{MP}) to the same data sets. Compute FL_E for onset and coda pairs and compare FL_E with FL_{MP} ➔ to improve the comparability of the results

- Perform statistical analysis ➔ to disentangle the relationship between the effect of position, phonological complexity and morphological typology

Thank you!