Cross-linguistic investigations of the organizational scales in phonological systems

Christophe Coupé, Egidio Marsico ,
Yooon Mi Oh & François Pellegrino

Laboratoire Dynamique du Langage
Université de Lyon & CNRS, France

Issue

How do the phonological units of a given language carry the weight of differentiating the words of its lexicon?

- Segments
- Syllables
- Features

Approach (I)

The notion of Functional Load (FL)

i. Relates to the role a phonological contrast plays in keeping words distinct in a given language (e.g. Trubeztkoy, 1939)

ii. Has been considered with respect to language evolution (Martinet, 1977; King, 1967; Surendran & Niyogi, 2003), language acquisition (Van Severen et al., 2012) etc.

iii. Has been seen as a useful, supplement to “standard” phonological descriptions.

iv. Mostly centered on phonemes, but some investigations into the FL of features (Surendran & Niyogi, 2003)

Approach (II)

A quantitative approach to FL to shed light on the organization of phonological systems

A cross-linguistic, corpus-based approach

Vowels, consonants (but also tones/stress)

Material: digital lexicons

(with frequency of use of forms)

Language	ISO 639-3 Code	Source
Cantonese	YUE	A linguistic corpus of mid-20th century Hong Kong Cantonese (Research Centre on Linguistics and Language Information Sciences, 2013)
English	ENG	WebCelex (Max Planck Institute for Psycholinguistics, 2013, 2014)
Japanese	JPN	The corpus of spontaneous Japanese (NINJAL, 2011)
Korean	KOR	(Leipzig corpora collection)
Mandarin	CMN	Chinese Internet Corpus (Sharoff et al, 2006)
German	DEU	WebCelex (Max Planck Institute for Psycholinguistics, 2013, 2014)
Swahili	SWH	Helsinki corpus of Swahili (Gelas, Besacier, & Pellegrino, 2012)
Italian	ITA	PAISÀ Corpus (Lyding et al., 2014)
French	FRA	Lexique 3.80 (New et al., 2001)

Simple vs. complex syllabic structures (Maddieson et al., 2013), different morphological types

20,000 most frequent words (inflected forms) considered, except for Cantonese (5,000 forms) & Italian (15,788 forms)

Methodology (I)

Carter (1987)'s quantitative definition of FL

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

FL of a contrast x/y = **quantification of the perturbation induced by merging *x* and *y* in terms of increase of homophony and of changes in the distribution of word frequencies**

FL(x,y): relative difference in entropy between the observed state *L* and a fictive state *Lxy** in which the contrast is neutralized

$$FL(x,y) = \frac{H(L) - H(L^*_{xy})}{H(L)} \qquad H(L) = -\sum_{i=1}^{N_L} p_{w_i}.log_2(p_{w_i})$$

Methodology (II)

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

Contrast a-i

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

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 = **31.8 %**

Inventory: /a i u p b l/

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

Phoneme /a/

From contrasts to units

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

FL(a) = ½ (FL(a-i)+FL(a-u)) = ½ (31.8+23.1)=27.45 %

First scale of organization: segments

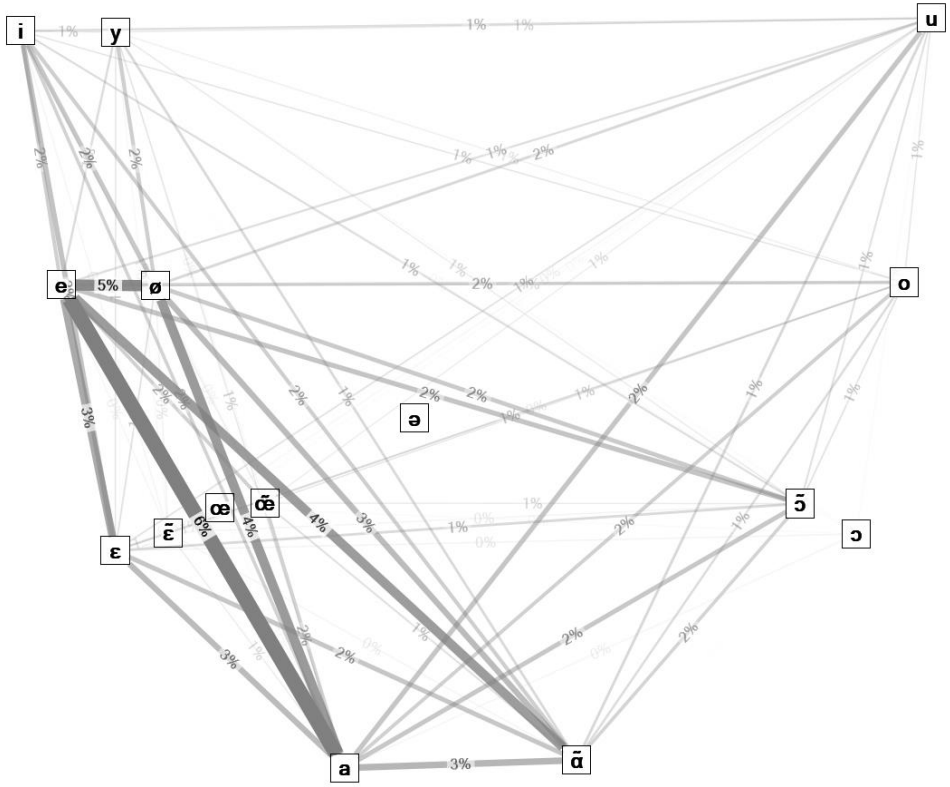
Language	ISO 639-3 Code	Phonological system	
Cantonese	YUE	V C T	13 19 6
English	ENG	V C S	22 28 2
Japanese	JPN	V C	10 16
Korean	KOR	V C	8 22
Mandarin	CMN	V C T	7 25 5
German	DEU	V C S	22 24 1
Swahili	SWH	V C	5 30
Italian	ITA	V C T	8 25 1
French	FRA	V C	15 21

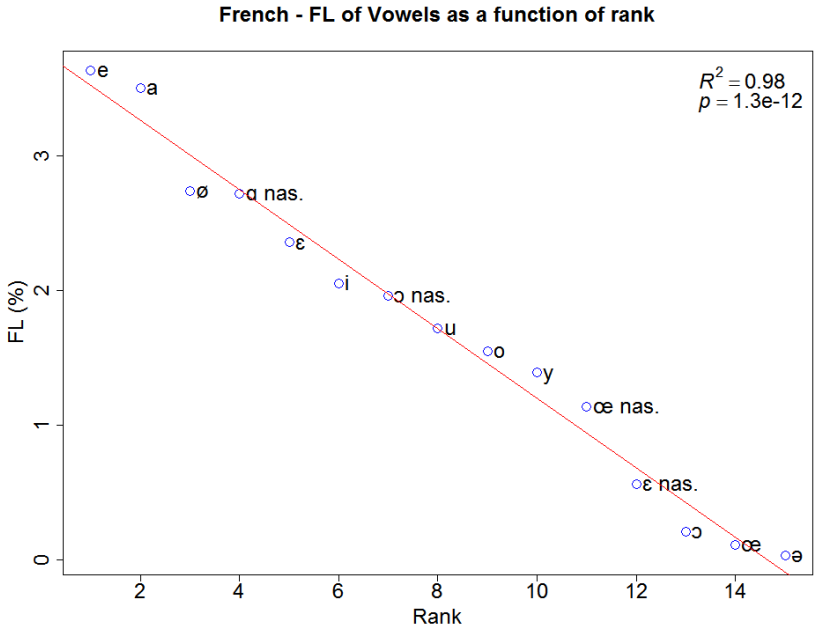
Vowels include diphthongs

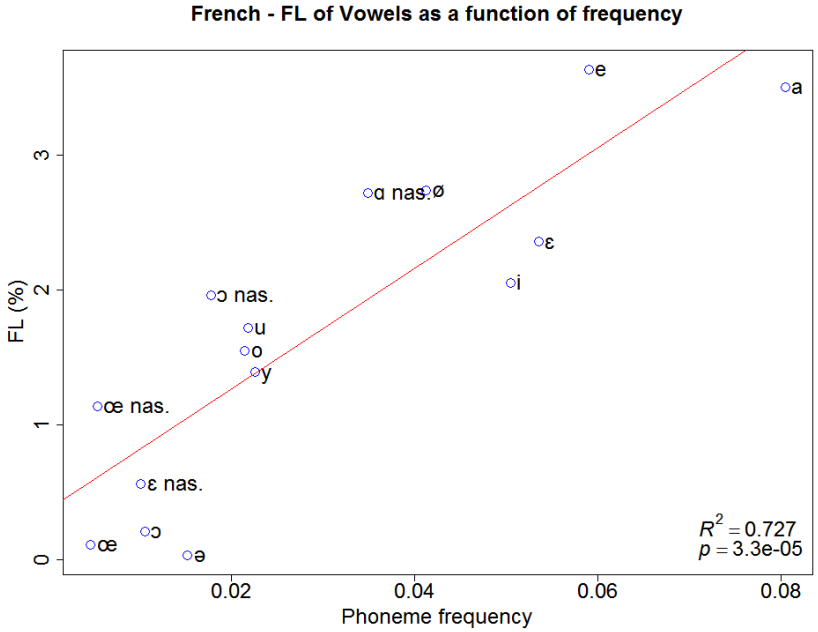
Languages differ in terms of the phonemes they use.

Some languages rely on much more vowels or consonants than others

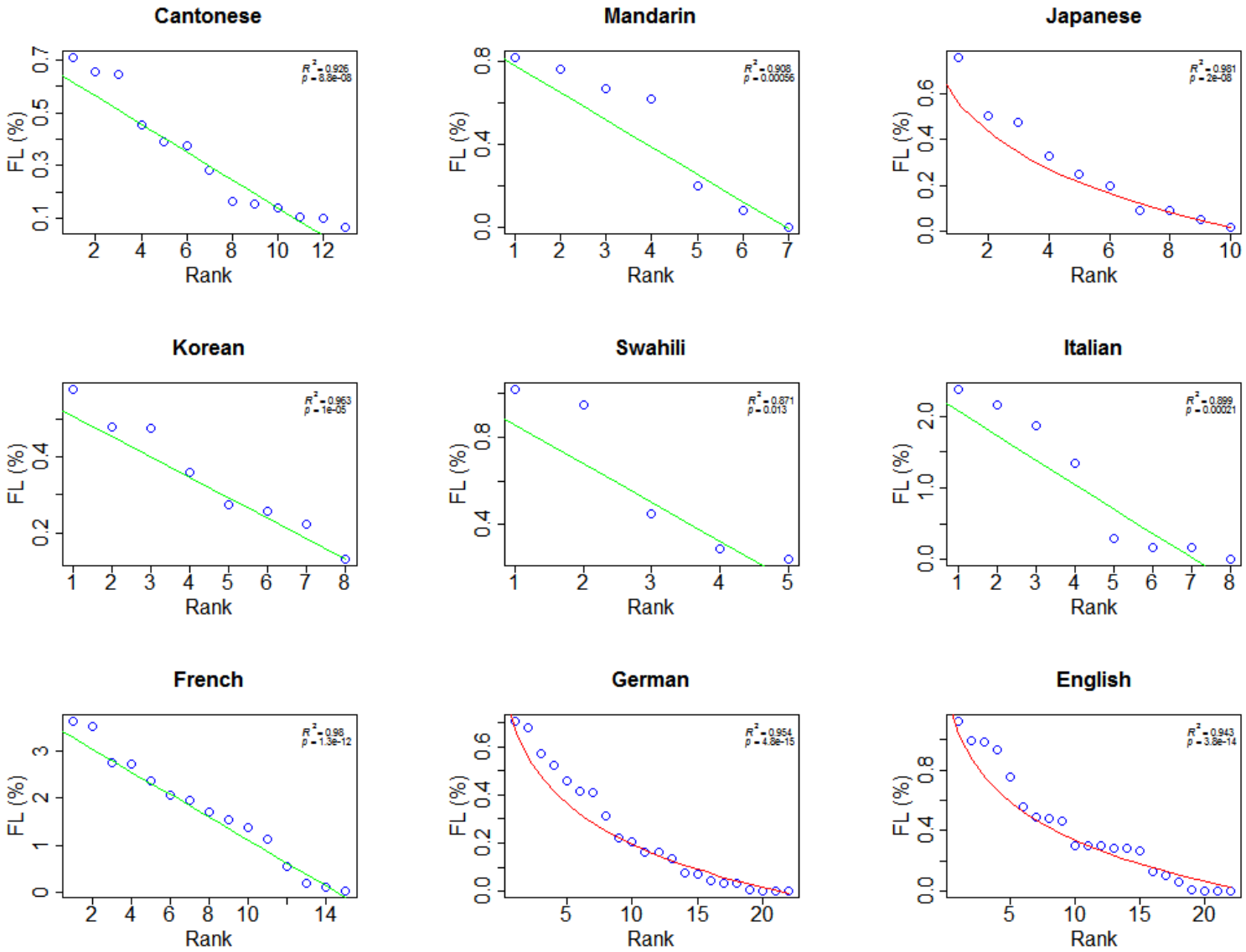
FL of vowels in French



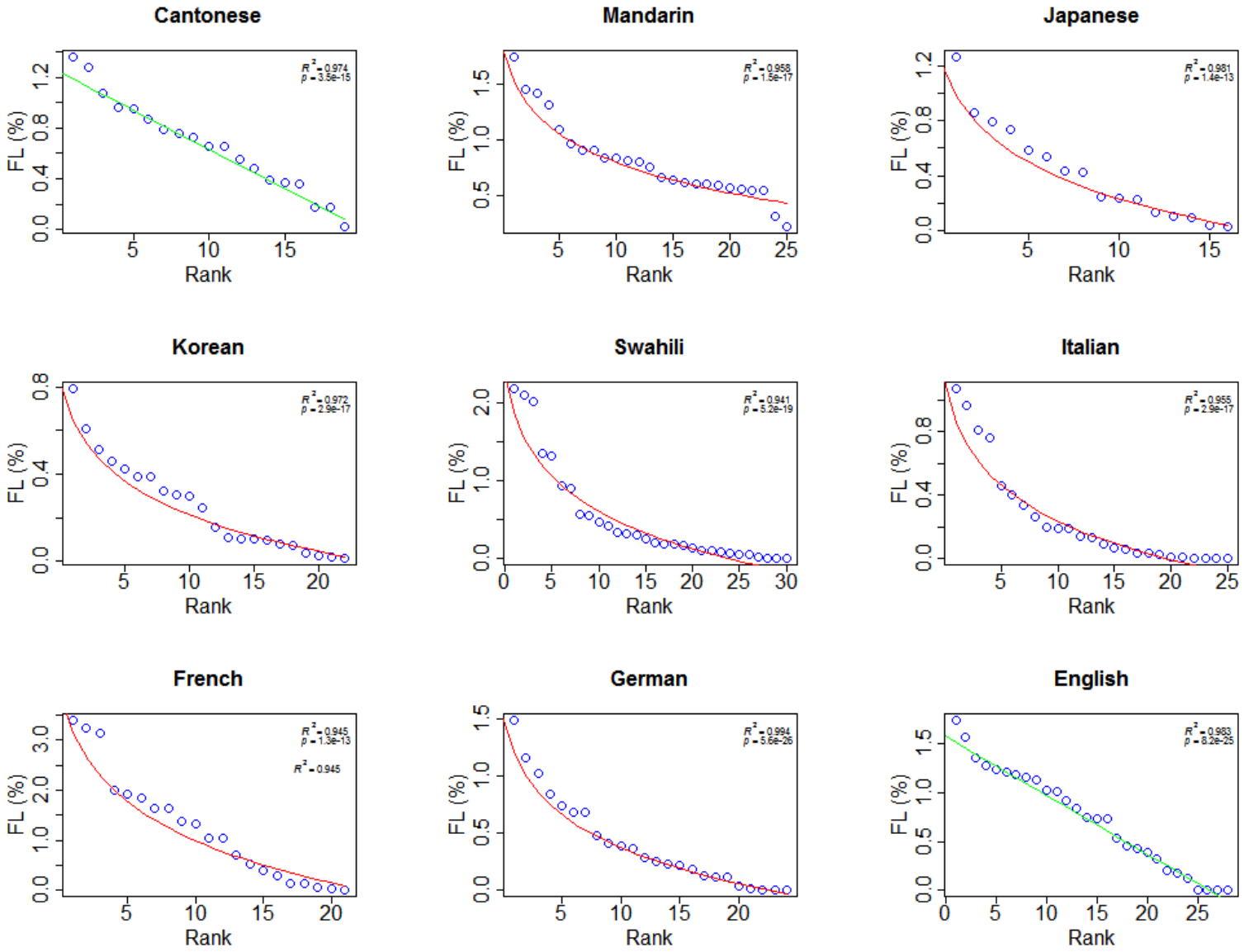




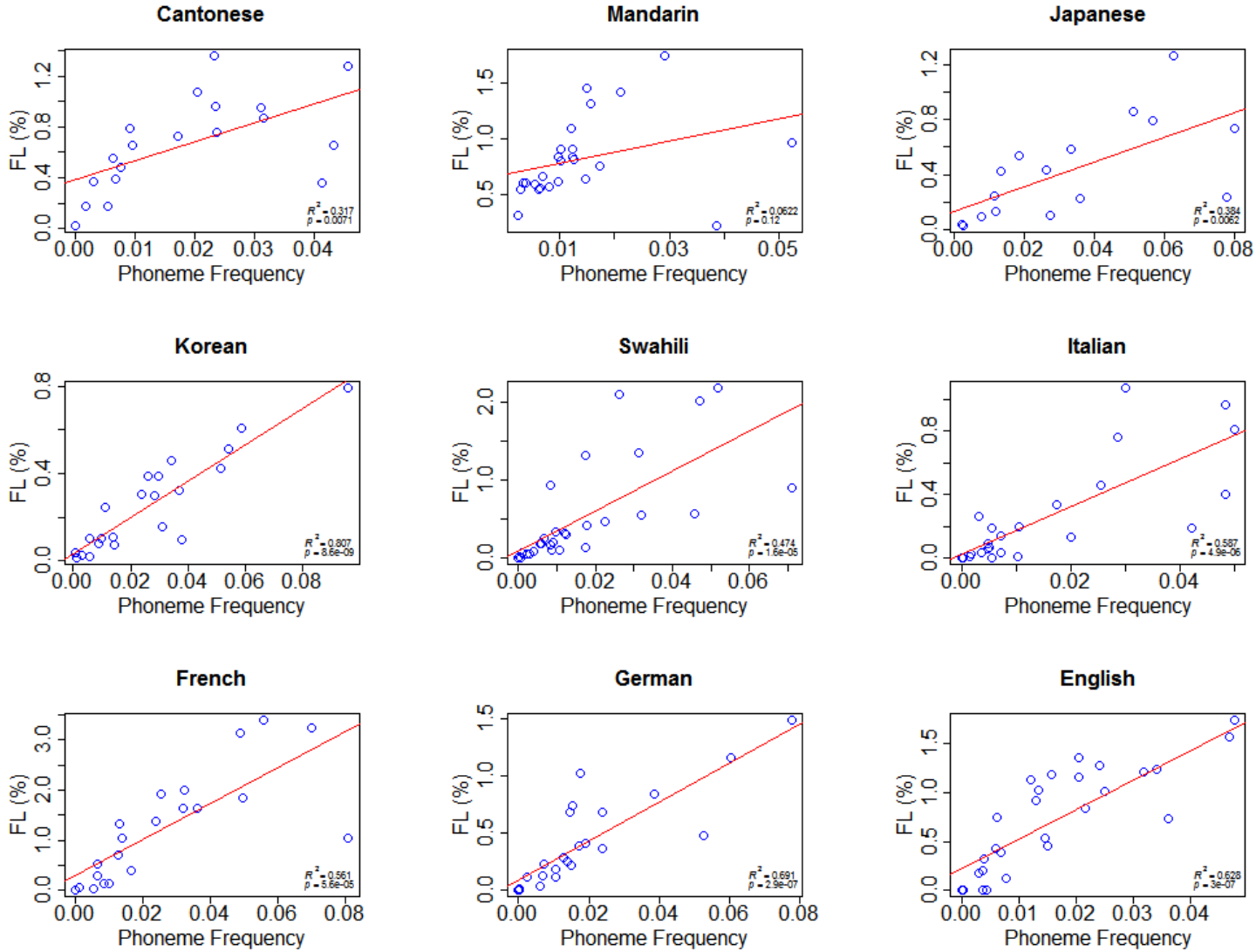
FL of vowels in 9 languages



FL of consonants in 9 languages



FL of consonants as a function of frequency



Discussion on segments

In all languages, some segments carry a heavier burden of differentiating the words than others

→ No principle of uniformity regarding functional load

Frequency is a partial predictor of FL, but other factors underlie the distinctive function within a system

Left for discussion time:

The segments with high FL differ widely from one language to another

There seem to be no *strong* cross-linguistic tendency.

Supra-segmental phonemes: the FL of tones is equivalent to the FL of vowels in tonal languages (yue & cmn)

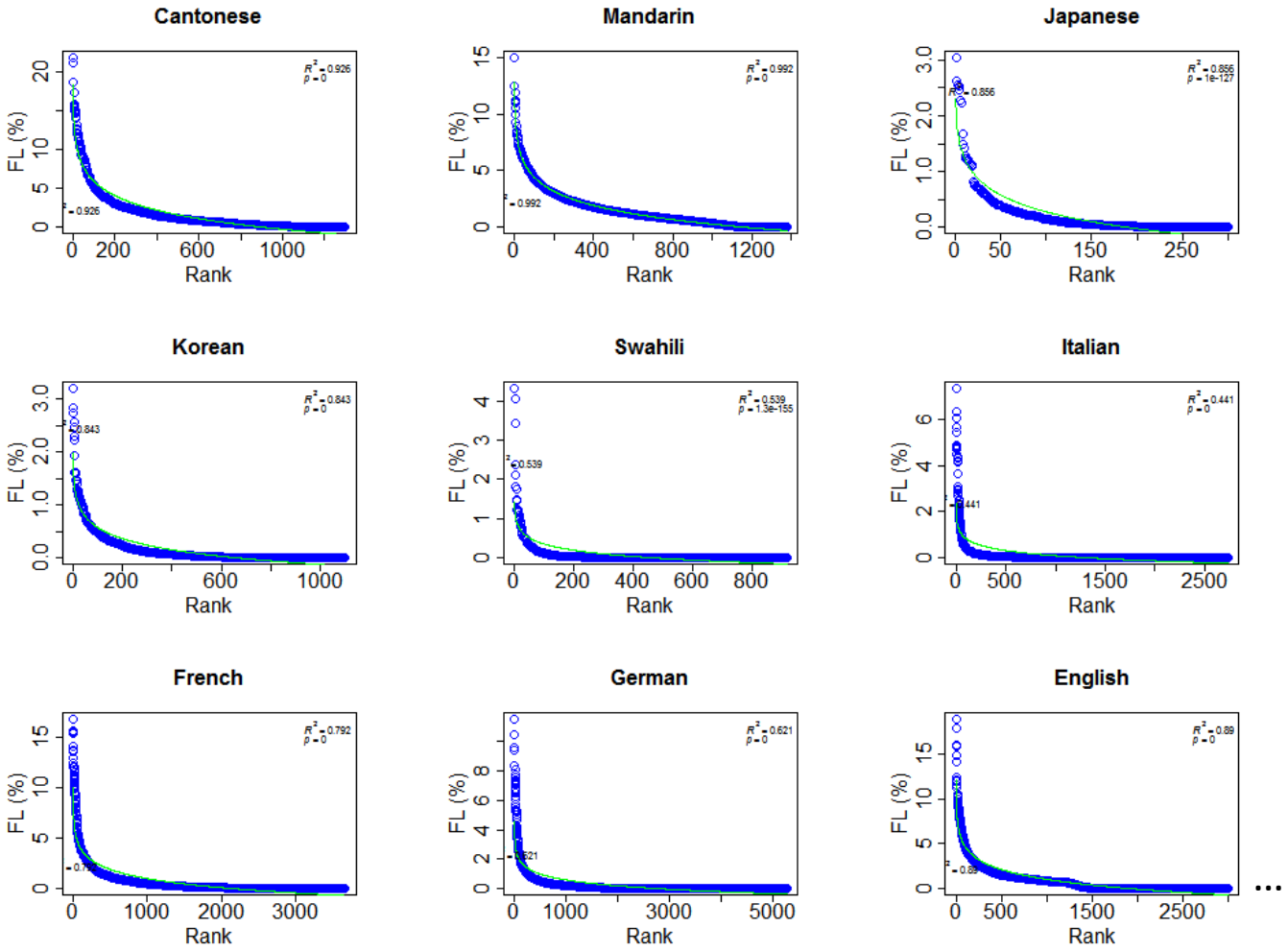
Second scale of organization: syllables

Language	ISO 639-3 Code	Phonological system		Number of different syllables
Cantonese	YUE	V C T	13 19 6	**1298**
English	ENG	V C S	22 28 2	**8298**
Japanese	JPN	V C	10 16	**300**
Korean	KOR	V C	8 22	**1100**
Mandarin	CMN	V C T	7 25 5	**1283**
German	DEU	V C S	22 24 1	**5256**
Swahili	SWH	V C	5 30	**914**
Italian	ITA	V C T	8 25 1	**2735**
French	FRA	V C	15 21	**3666**

Languages differ widely in terms of number of syllables

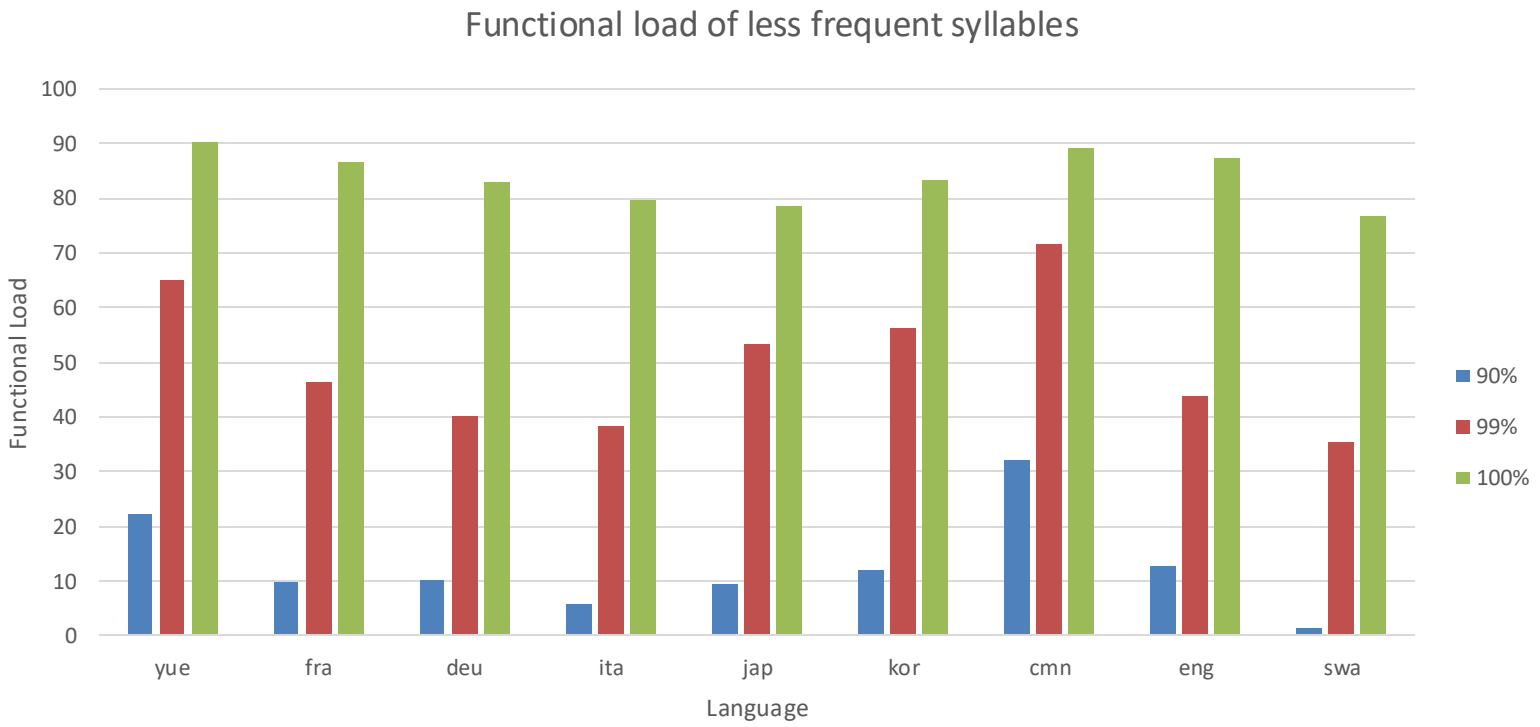
The FL of syllables can be computed in the same way they were for segments

FL of syllables according to rank



Relative to segments, the distribution of syllabic FL is more skewed (esp. in languages like **deu**, **ita** or **swa**): in each language, a large number of syllables have a very low FL.

FL of less frequent syllables



The green bars refer to a situation where words are only differentiated by their number of syllables, not by the nature of these syllables

Even when 90% of the (less frequent) syllables are merged into one, words can still be well differentiated thanks to their structure and to the most frequent syllables
Cantonese and Mandarin are more affected than other languages

Discussion on syllables

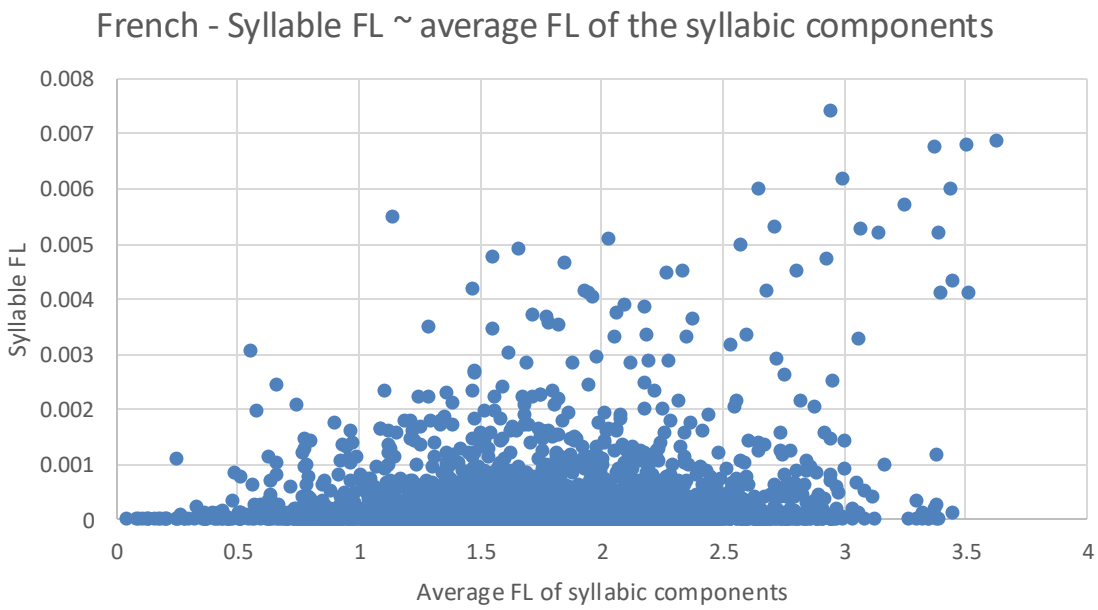
Non-linear and partial relationship between FL and frequency of use (sometimes strong, e.g. yue & cmn) (*not shown*)

Why do we have so many syllables, when most of them have very low FL?

Is this distribution, at the syllabic level, functional? or…

Does the FL of syllables derive from the FL of their components?

The FL of syllables does not correlate with either the average, the maximum value or the product of the FL of its components
Is there another relationship?



Third scale of organization: features

Language	ISO 639-3 Code	Phonological system		Number of features
Cantonese	YUE	V C	13 19	12 18
English	ENG	V C	22 28	27 19
Japanese	JPN	V C	10 16	10 15
Korean	KOR	V C	8 22	10 17
Mandarin	CMN	V C	7 25	11 19
German	DEU	V C	22 24	21 18
Swahili	SWH	V C	5 30	9 19
Italian	ITA	V C	8 25	10 18
French	FRA	V C	15 21	12 17

E.g.

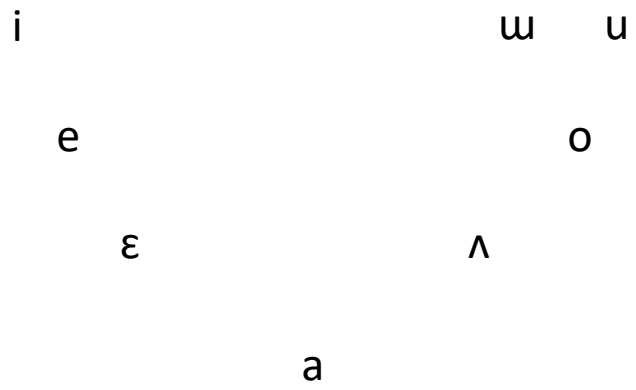
/i/: high front unrounded

/p/: bilabial voiceless stop

(Mostly articulatory) description of segments in terms of features based on UPSID (Maddieson & Precoda, 1990)

Obvious differences among languages. Some articulatory dimensions and specific features are always present, others are not.

How to compute the FL of features?



Vocalic inventory of Korean

Aperture: 3 sets of merges

ɛ, e, i
o, u
ʌ, ɯ

Anteriority: 2 set of merges

ɛ, ʌ
i, ɯ

E.g. To compute the FL of **aperture**, the initial lexicon is contrasted with a lexicon where 3 different sets of merges create homophony and modify the distribution of word frequencies.
Note: consistent combinations of features (e.g. front-unrounded) not considered here

FL of articulatory dimensions

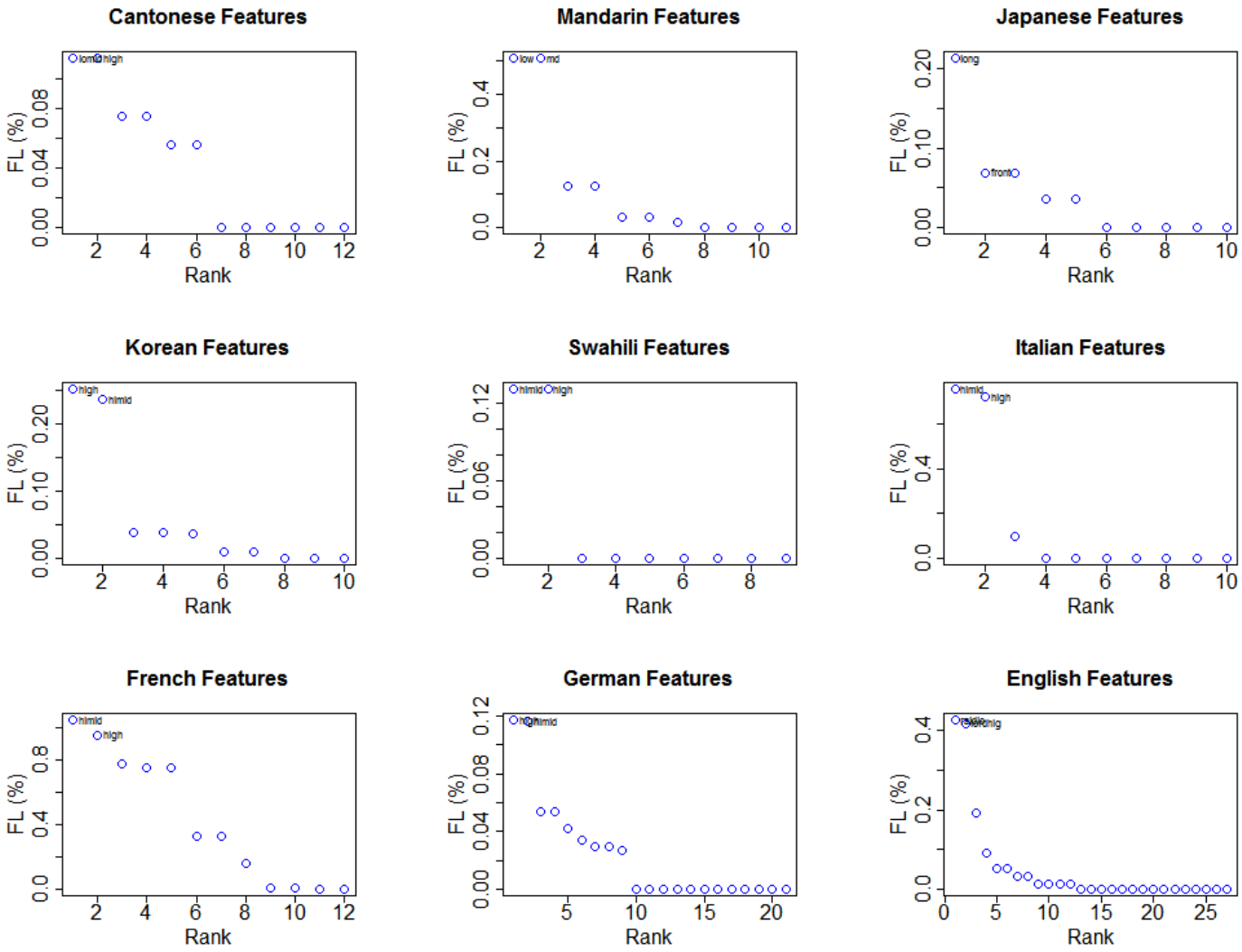
	V						C		
	Aperture	*Anteriority*	*Roundedness*	*Length*	*Aspiration*	*Nasalization*	*Place*	*Manner*	*Voicing*
yue	0,23	0,15	0,11		0,33		1,72	1,07	
cmn	1,02	0,06	0,25		0,67		2,07	1,04	0,18
jap	0,07	0,14		0,21			0,79	0,26	0,36
kor	0,53	0,08	0,02		0,01		0,67	1,01	0,05
swa	0,26						1,84	1,66	0,13
ita	1,62						0,22	2,03	0,12
fra	2,66	0,66	1,50			0,16	1,56	3,04	0,89
deu	0,31	0,11	0,06	0,03			0,91	2,72	0,11
eng	1,19	0,11					0,99	2,45	0,59

	Primary articulatory dimension with the hightest FL
	Primary articulatory dimension with the 2nd hightest FL
	Primary articulatory dimension with the 3rd hightest FL

Regarding vowels and primary articulatory dimensions, **aperture** carries the heaviest load in 8 of the 9 languages. Secondary features can have a high / the highest FL.

Regarding consonants, languages seem to choose **either place or manner** as the primary way to differentiate between words. **Voicing** always comes after except in Japanese.

FL of vocalic features according to rank



The distribution is never uniform.

Most often, 1 or 2 features have a higher FL than others (but consistent combinations)

Discussion & perspectives

Cross-linguistic tendencies:

Whatever the organizational scale, phonetic units do not evenly carry the same FL. A few units carry a heavy load, while most others only carry a very light load.

Partial explanations can be provided at each level : morphosyntactic rules, ease of production versus sufficient perceptive contrast etc. (language-specific or not)

Cross-linguistic diversity:

Languages widely differ in the units they shoulder with the task of keeping words distinct

Are there 'vertical' integrative processes taking place in addition to 'horizontal' constraints?

The FL of syllables does not seem to derive straightly from the FL of their components

Perspectives:

Investigate whether the FL of segments derive from the FL of their features

Consider pairs of features in addition to single features.

Thank you for your attention

Yoonmi Oh

Egidio Marsico

François Pellegrino

Oh, Y., Pellegrino, F., Coupé, C. & Marsico, E., 2013. " Cross-language Comparison of Functional Load for Vowels, Consonants, and Tones". Proc. of Interspeech 2013, Lyon, France, 25-29 August.

Oh, Y., Coupé, C., Marsico, E. & Pellegrino, F. (accepted, to appear). "Bridging Phonological System and Lexicon: Insights from a Corpus Study of Functional Load". *Journal of Phonetics*.

Funding: LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon - Program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French government

ASLAN Advanced Studies on Language complexity
UNIVERSITE DE LYON