Functional Load and organization scales in phonological systems

Christophe Coupé, Yoon-Mi Oh, Egidio Marsico & François Pellegrino
"Dynamique Du Langage" Laboratory
CNRS & University of Lyon

Structural Complexity in Natural Language(s) (SCNL)
International Workshop – 30-31 May 2016, Paris, France

ASLAN
UNIVERSITE DE LYON

DDL

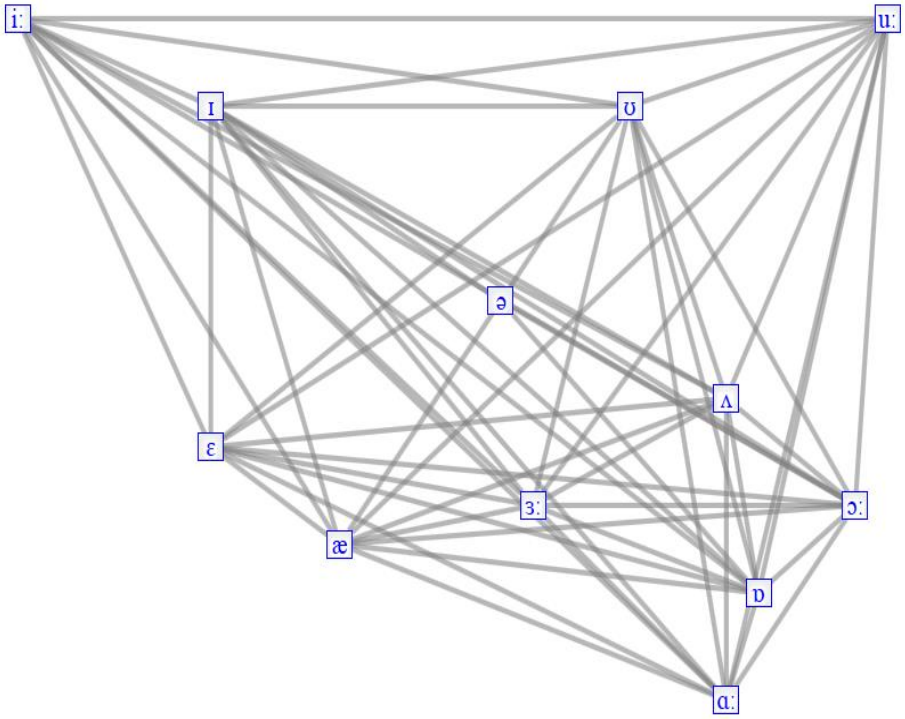
‘The function of a phonemic system
is to keep the utterances of a language apart.

Some contrasts between the phonemes in a system apparently do more of this job than others.’

Charles F. Hockett (1966)

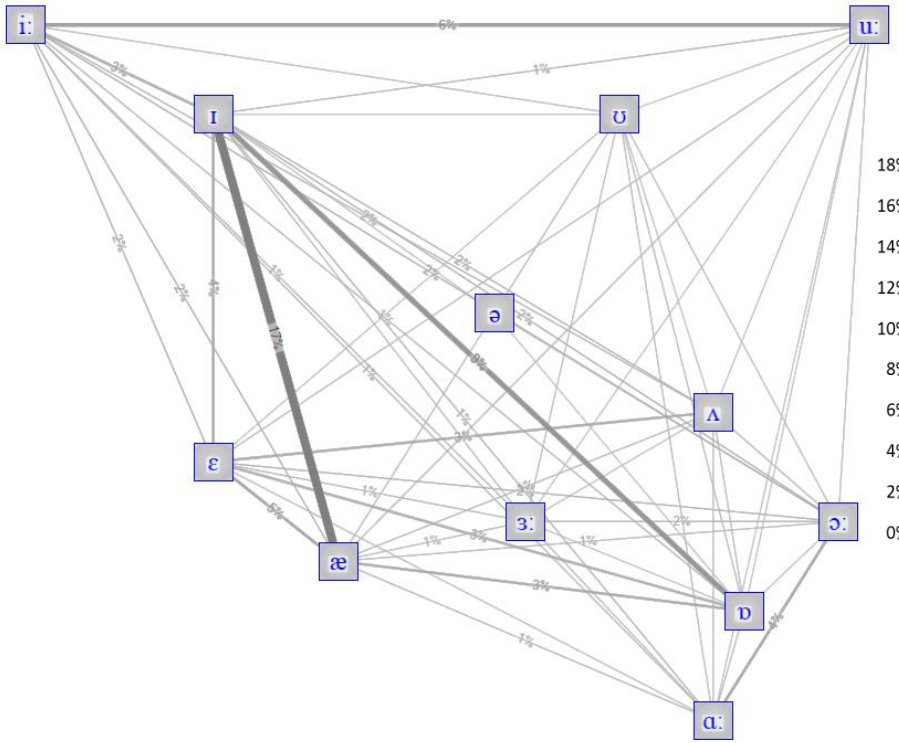
Vowel system as an (organized) set of vowel segments

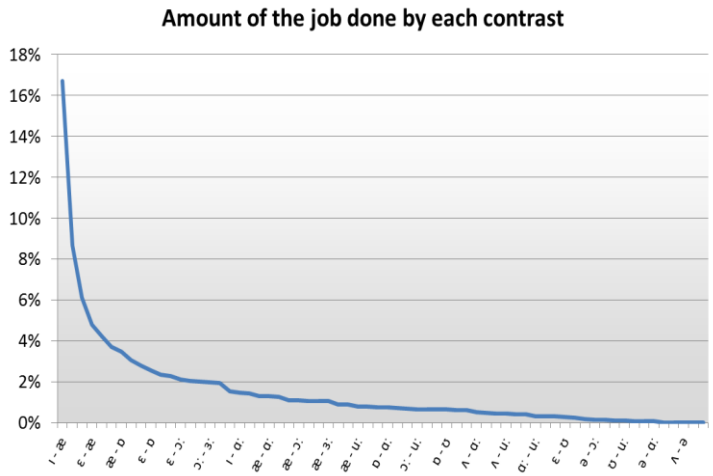
Vowel contrasts as a fully-connected graph



Equal Thickness = equal amount of the job done by each contrast

But some contrasts actually do more of the job…



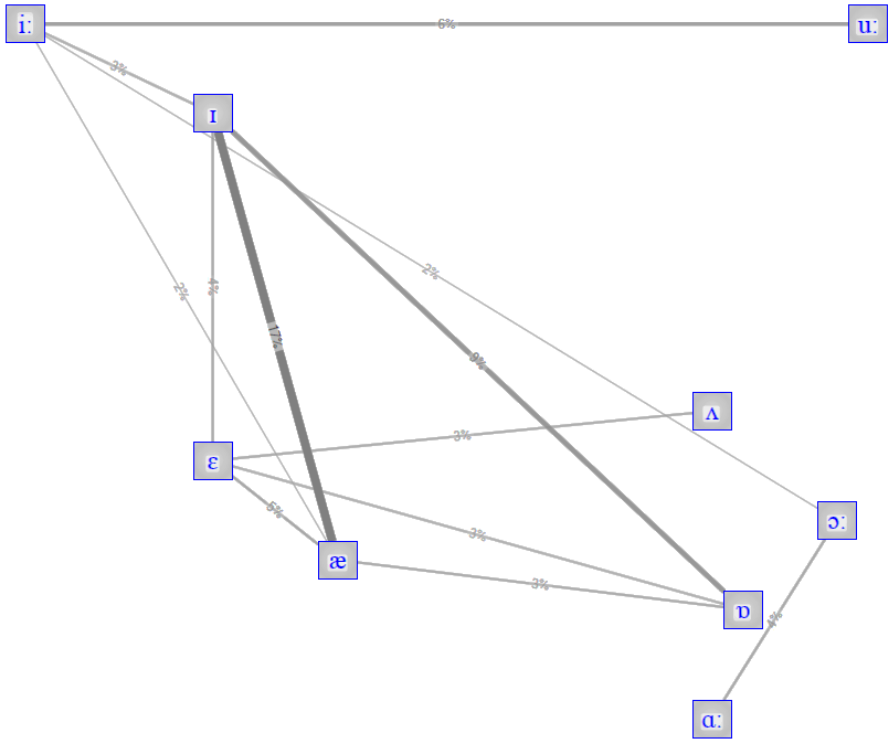


British English, CELEX database

Quiz: /v-ʒ/ contrast in English lexicon?

Thickness illustrates relative amount of the job done by each contrast

- Some segments are involved in none of the major contrasts… Are they nevertheless useful?



Less “important” contrasts and vowels erased

Functional Load and organization scales in phonological systems

Christophe Coupé, Yoon-Mi Oh, Egidio Marsico & François Pellegrino
"Dynamique Du Langage" Laboratory
CNRS & University of Lyon

Structural Complexity in Natural Language(s) (SCNL)
International Workshop – 30-31 May 2016, Paris, France

ASLAN
UNIVERSITE DE LYON

DDL

Christophe COUPÉ

Yoon Mi OH

Egidio MARSICO

Funding: LABEX ASLAN (ANR-10-LABX-0081),
University of Lyon, French Program "Investissements d'Avenir" (ANR-11-IDEX-0007)

LABORATOIRE DYNAMIQUE DU LANGAGE

ASLAN Advanced Studies on Language complexity
UNIVERSITE DE LYON

General framework

- **Complexity of phonological systems**

  ➔ Information as a link between description and usage

- **Phonological systems as complex adaptive systems**

  ➔ Search for new emergent regularities and trends

Research questions

How is (lexical) information distributed in phonological systems?

Could it change our view on phonological systems in a typological perspective ?

Does it tell us something interesting on languages (and speakers) ?

FORTHCOMING IN 2016…

Complexity in Language:
Developmental and Evolutionary Perspectives

edited by

Salikoko S. Mufwene
Christophe Coupé,
& François Pellegrino

ASLAN
UNIVERSITE DE LYON

DDL

Position

- Typological view → Cross-linguistic perspective
  - ✓ More than a dozen languages considered
- Systemic approach
  - ✓ Based on the relationship among phonological units, extracted from the lexicon
- Information-theory approach
  - ✓ Entropy-based estimation of *Functional Load*
- Multi-scale approach
  - ✓ Features Segments Syllables Subsystems

Overview

- The notion of Functional Load
- Methodology & Data
- Results
- Conclusion

THE NOTION OF FUNCTIONAL LOAD (FL)

The origins

Cercle Linguistique de Prague

- ✓ Travaux du CLP 1 (1929) *Tâches fondamentales de la phonologie synchronique*.
  - "(…) Il faut également étudier la charge fonctionnelle des divers phonèmes et combinaisons de phonèmes dans une langue donnée".
- ✓ Travaux du CLP 4 (1931)
  - 'Rendement fonctionnel: Degré d'utilisation d'une opposition phonologique pour la différenciation des diverses significations des mots dans une langue donnée'. (probable author: Mathesius)

Trubetzkoy (1939)

- 'it is also possible to determine (…) the extent to which the individual phonological oppositions are utilized distinctively (their functional load) as well as the average load of the phonemes in general. It develops that there are "economical" and "wasteful" languages in this respect (…).' (Trubeztkoy, Principles, 1969:268).

The origins (cont'd)

Martinet, (since 1933)

- ✓ Link between *Functional Load* (rendement fonctionnel) and *Diachrony*

  '(…) une opposition phonologique qui sert à maintenir distincts des centaines de mots parmi les plus fréquents et les plus utiles n'opposera-t-elle pas une résistance plus efficace à l'élimination que celle qui ne rend de service que dans un très petit nombre de cas ?' (1955).

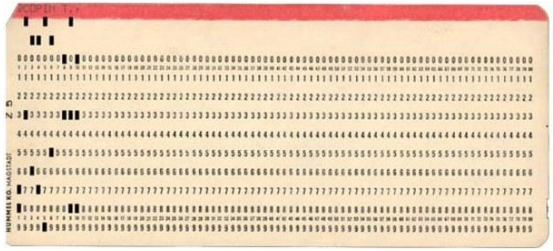
- ✓ See also *Functional burdening* (Twaddell, 1935)

Comments on these seminal works

- ✓ Strong intuitions, but lack of data and of mathematical concepts to test them
- ✓ Functional Load is nevertheless cited
  - In passing, as a valuable addition to any description of a phonological system (Hockett, 1955)
  - As a relevant factor of resistance/propensity to sound change (Martinet, 1955; Hoenigswald, 1960)
- ✓ The diffusion of Shannon's Communication theory (aka Information theory) will provide conceptual tools to go beyond mere intuitions

Some quantitative developments (but limited results)

- Hockett (1955, 1961/1966)
  - ✓ Inspired by Martinet
  - ✓ Methodological proposal: FL = loss of entropy under the hypothesis of a phoneme coalescence
  - ✓ But no implementation and assessment with real data
- Kučera (1963)
  - ✓ Corpus-based comparison of Russian and Czech (IBM 7070)
  - ✓ Phonemic and syllabic inventory entropies, FL of features



- King (1967)
  - ✓ First in-depth and corpus-based evaluation of the role of FL in sound change (IBM 7040)
  - ✓ Martinet's hypotheses tested against 4 historical changes (in the Germanic family)
  - ➔ "Functional load, if it is a factor in sound change at all, is one of the least important of those we know anything about (…)".
- See also Wang (1967)

To the dead-concept rubbish can?

A new century, a new dawn for FL…

- Initiative: D. Surendran and colleagues
- First "modern times" corpus studies
  - ✓ FL estimation for several types of contrasts (features, phonemes, tones)
    - "The Functional Load of tone in Mandarin is as high as that of vowels"
  - ✓ Testing of intuitions in historical linguistics and language acquisition
    - Historical linguistics: n-l merger in Cantonese: high FL (rejects Martinet's hypothesis)
    - Historical linguistics: No correlation between ease of articulation and FL
    - Language acquisition: But FL pretty well predicts the age of acquisition of consonants for American English children (but no significant correlation is observed in Cantonese)

Stokes, S. and D. Surendran. 2005. Articulatory complexity, ambient frequency and functional load as predictors of consonant development in children, *Journal of Speech and Hearing Research* 48(3)
Surendran, D. and G.-A. Levow. 2004. The Functional Load of Tone in Mandarin is as High as that of Vowels. in *Proc. of Speech Prosody 2004*, Japan.
Surendran, D. and P. Niyogi. 2003. *Measuring the Usefulness (Functional Load) of Phonological Contrasts*. Technical Report TR-2003-12., Department of Computer Science, University of Chicago.

Recent works

- Van Severen, L., Gillis, J. J., Molemans, I., Van Den Berg, R., De Maeyer, S., & Gillis, S.
  - ✓ The relation between order of acquisition, segmental frequency and function: the case of word-initial consonants in Dutch. *Journal of Child Language*, 40(4). 2013.
- Wedel, A., Kaplan, A., & Jackson, S.
  - ✓ High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2). 2013.
- Our group
  - ✓ Pellegrino, F., Marsico, E. & Coupé, C., (2012), "Vowel inventories revisited: the functional load of vowel contrasts", *86th Annual Meeting – Linguistic Society of America* , Portland, USA.
  - ✓ Oh, Y.M., Pellegrino, F., Coupé, C. & Marsico, E., (2013). "Cross-language comparison of functional load for vowels, consonants, and tones". *Proc. Interspeech 2013*, Lyon, France.
  - ✓ 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.

METHODOLOGY & DATA

Methodology – Notion of entropy

- Mathematical theory of communication (Shannon, 1948)
  - ✓ A theory of communication (= information transmission)
  - ✓ Quantification of information, entropy, channel capacity and redundancy
  - ✓ Food for linguistic thought…
    - Hockett's review of Shannon's theory (1953 – more than 20 pages!)
    - Cherry, Halle and Jakobson (1953)
- Considering that language *L* is a source of linguistic sequences composed of units (*w*) from a finite set (N_L)
- Entropy *H*(*L*) = Average quantity of information per unit

$$H(L) = -\sum_{i=1}^{N_L} p_{w_i} \log_2 \left(p_{w_i} \right)$$

  - ✓ Easy to estimate from the set of units and their probabilities
  - ✓ Probabilities p_{w_i} estimated by their frequency in a relevant corpus
  - ✓ Units may be words, syllables, phonemes, etc.

ASLAN UNIVERSITE DE LYON DDL

Methodology – FL estimation

Quantitative entropy-based definition of FL

- ✓ Following Hockett (1966) & Carter (1967)
- ✓ Language *L* considered as a source of sequences of independent words w_i taken from a 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 L^*_{xy} in which the contrast is neutralized

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

ASLAN UNIVERSITE DE LYON

DDL

Methodology – Toy Language

Observed Lexicon

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

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 = **31.8 %**

FL(a-u) = **23.1 %**

FL(i-u) = **21.0 %**

Inventory: /a i u p b l/

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

Phoneme /a/

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

FL_V=61%

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

ASLAN UNIVERSITE DE LYON

DDL

Data overview: Recipe

- For each language
- Fetch a large written or transcribed oral corpus
- Extract the lexicon with frequencies of occurrence for each word
- Syllabify and Phonetize the lexicon
- Define the contrasts of interest and evaluate their functional loads

- Icing on the cake (optional, unfortunately)
  - ✓ Lexicons already processed
  - ✓ Morphosyntactic information available (Part of speech tagging, Lemmatization)

ASLAN UNIVERSITE DE LYON

DDL

Material

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	Gelas, Besacier, & Pellegrino, (2012)
Italian	ITA	PAISÀ Corpus (Lyding et al., 2014)
French	FRA	Lexique 3.80 (New et al., 2001)

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

ASLAN UNIVERSITE DE LYON

DDL

RESULTS

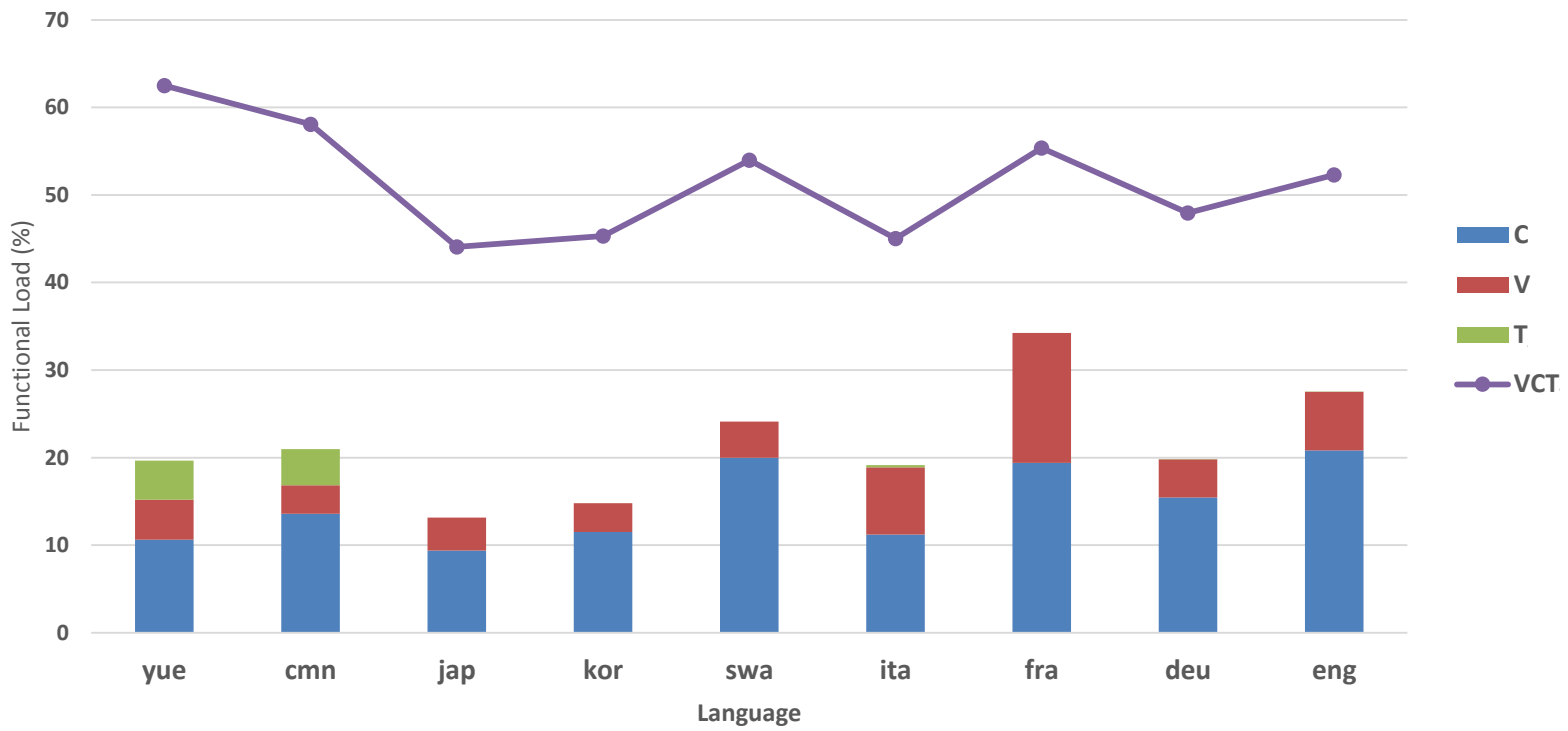
Subsystems

Segments

Syllables

Features

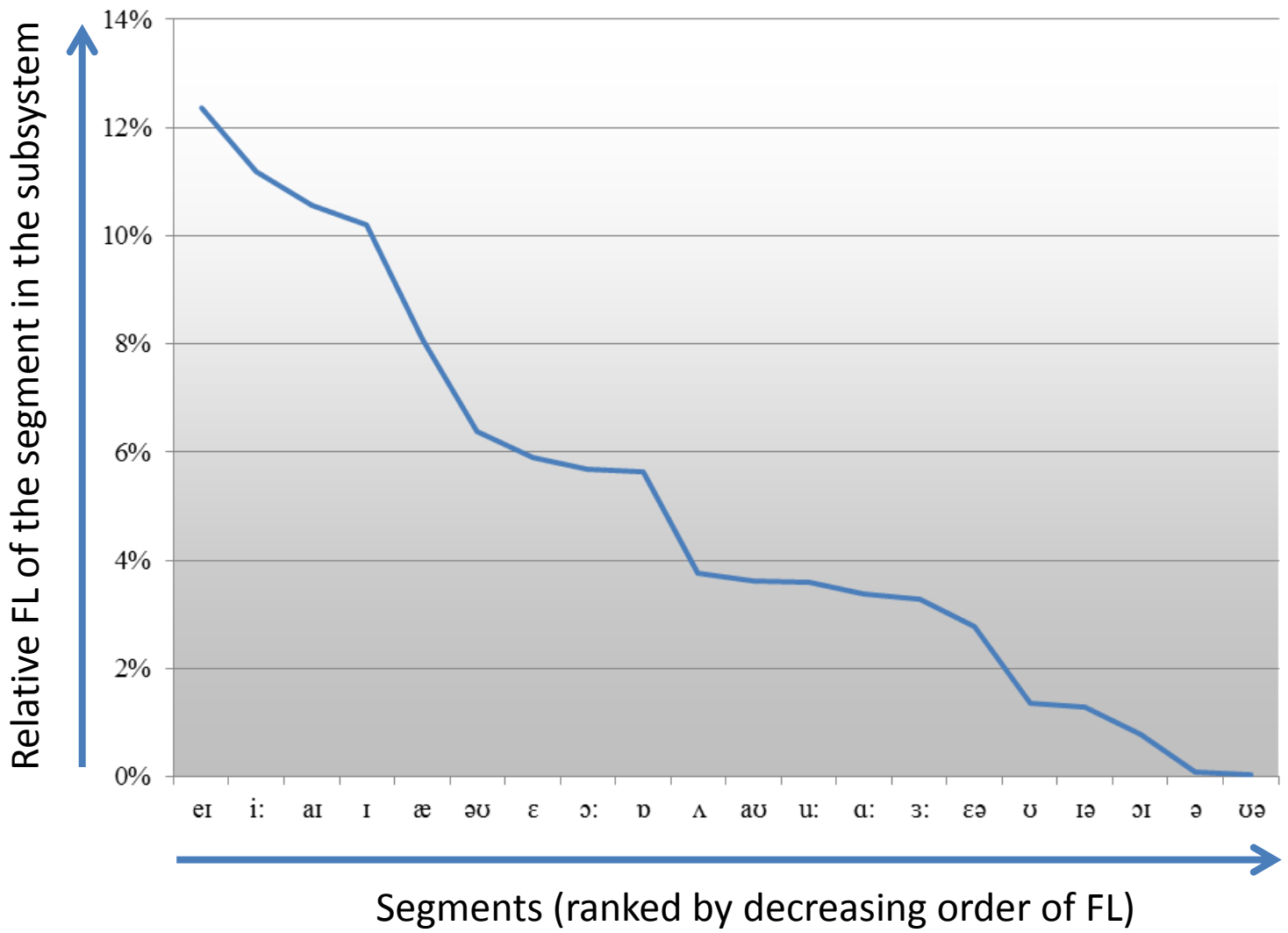
FL OF VOWELS, CONSONANTS AND TONES



- Variation in phonological FL
- In Mandarin and Cantonese, $FL_V \approx FL_T$
- High FL of vowels in French (and Italian)

ASLAN
UNIVERSITE DE LYON
DDL

Result Illustration



Most "natural" scale of organization: segments

Language	ISO 639-3 Code	Phonological system size	
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 S	8 25 1
French	FRA	V C	15 21

Oh, Coupé, Marsico, & Pellegrino (2015). *J. Pho.*

⌦ Methodological details

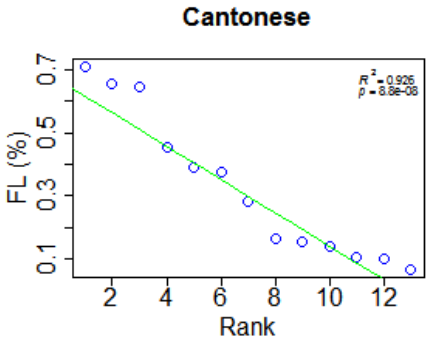
⌦ Discussion

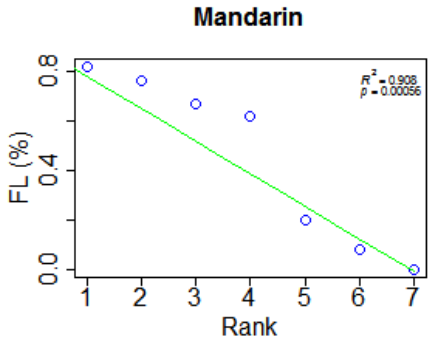
- Lemmas vs inflected wordforms
- Types vs. Tokens
- Consonantal bias hypothesis (Nespor, Peña, & Mehler, 2003)

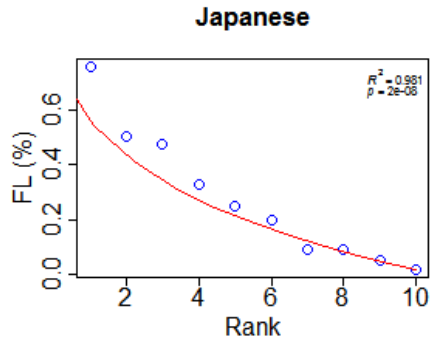
ASLAN UNIVERSITE DE LYON

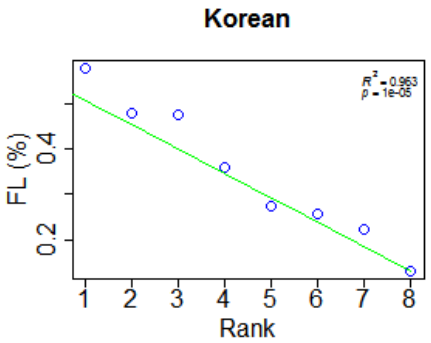
DDL

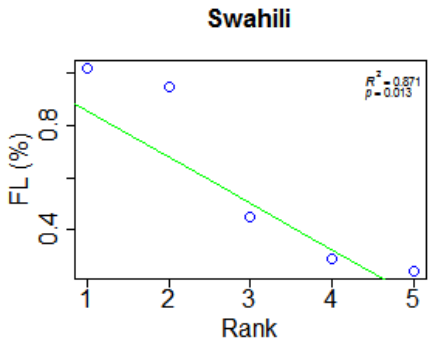
VOWEL FL

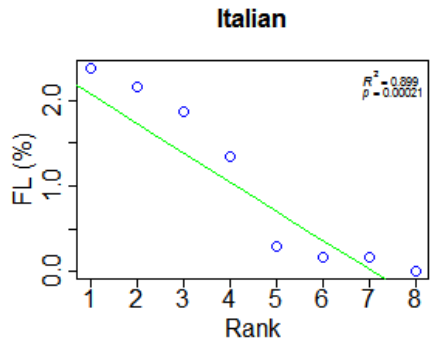


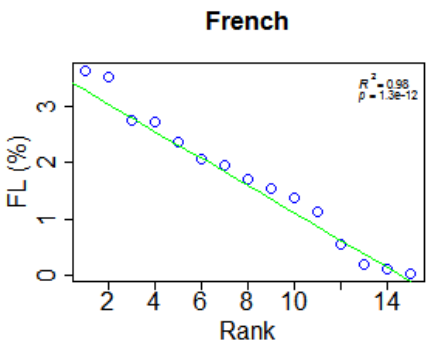


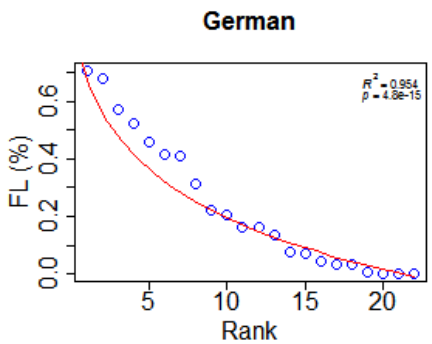


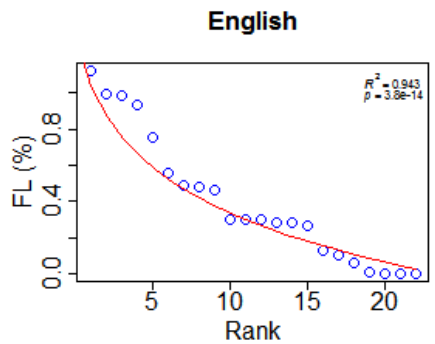








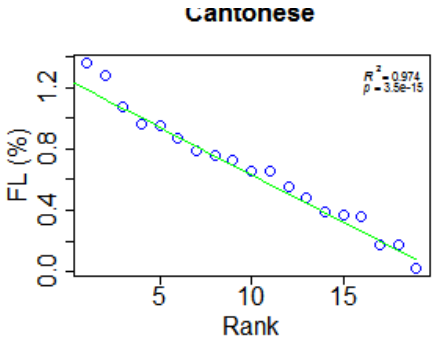


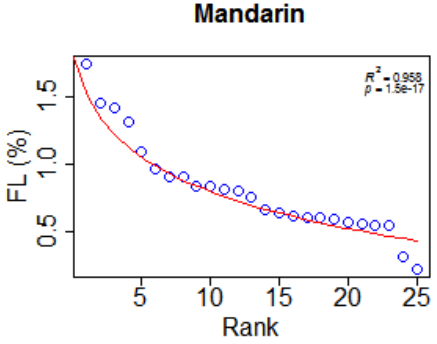


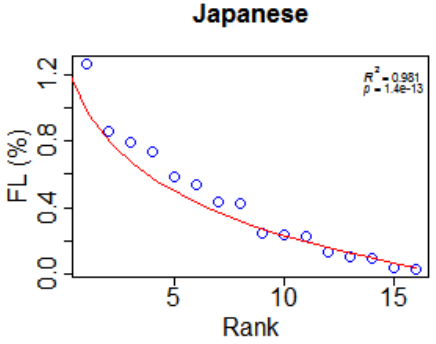
ASLAN
UNIVERSITE DE LYON

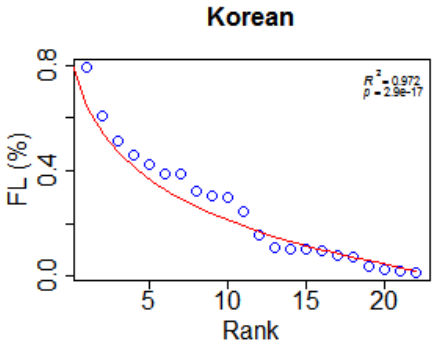
DDL

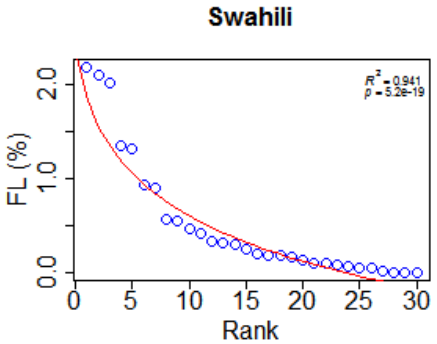
CONSONANT FL

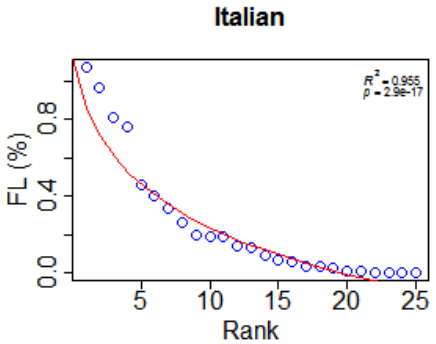


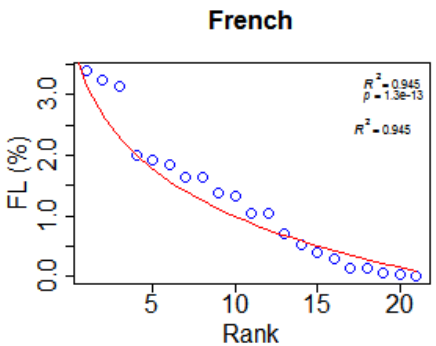


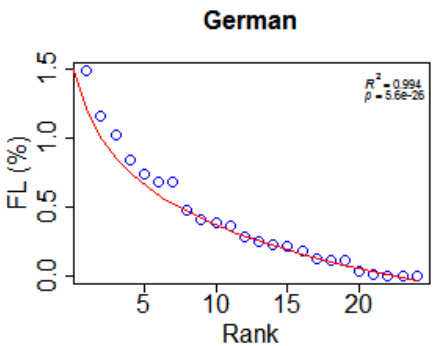


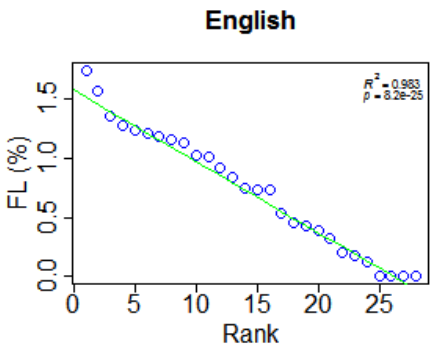












ASLAN
UNIVERSITE DE LYON

DDL

Syllabic scale

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

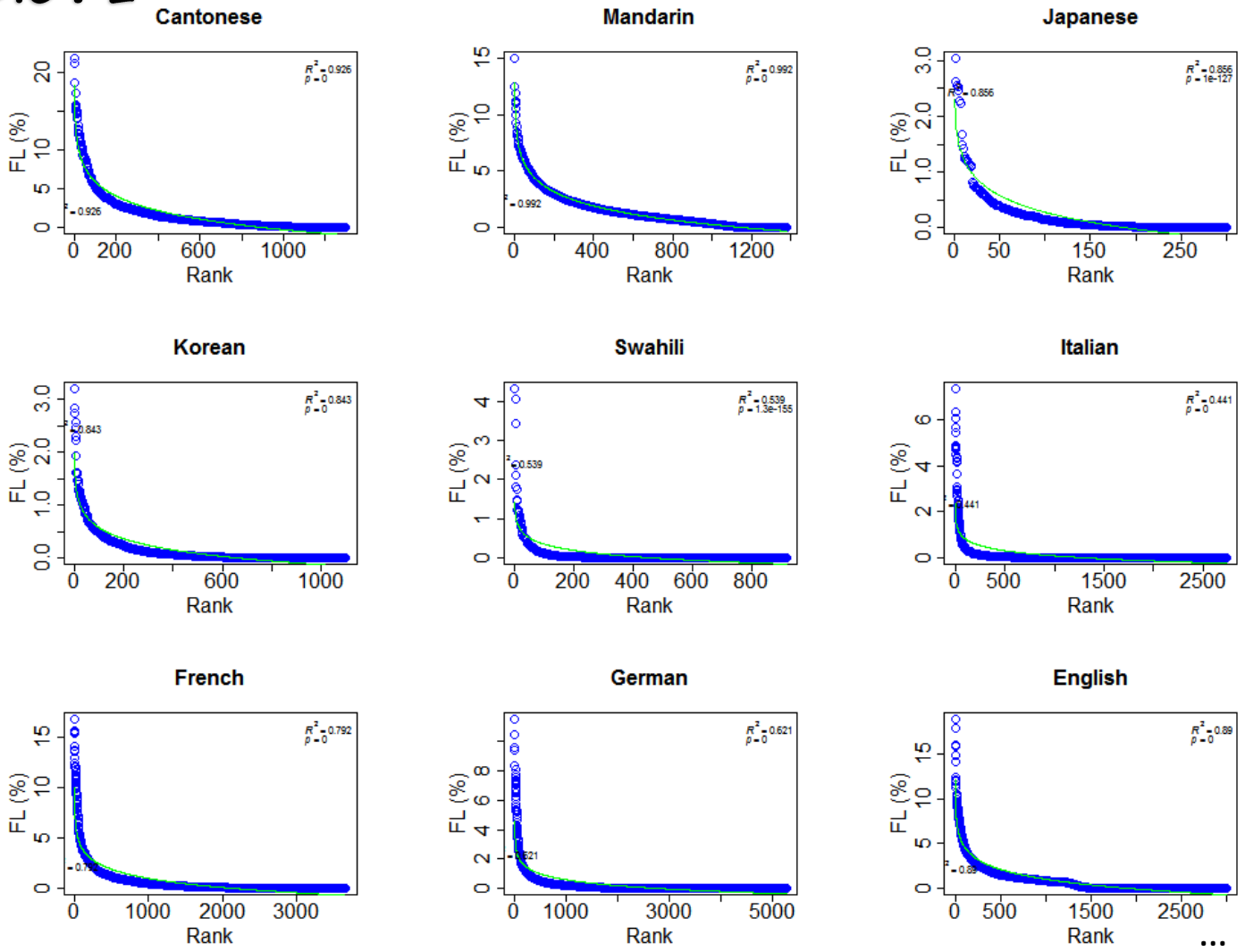
⌦ Languages widely differ in terms of syllable inventory

⌦ Syllable FLs can be computed in the same way they were for segments

ASLAN UNIVERSITE DE LYON

DDL

Syllabic FL



Distribution of syllabic FL highly skewed: large numbers of syllables have very low FLs

ASLAN UNIVERSITE DE LYON

DDL

Feature scale

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

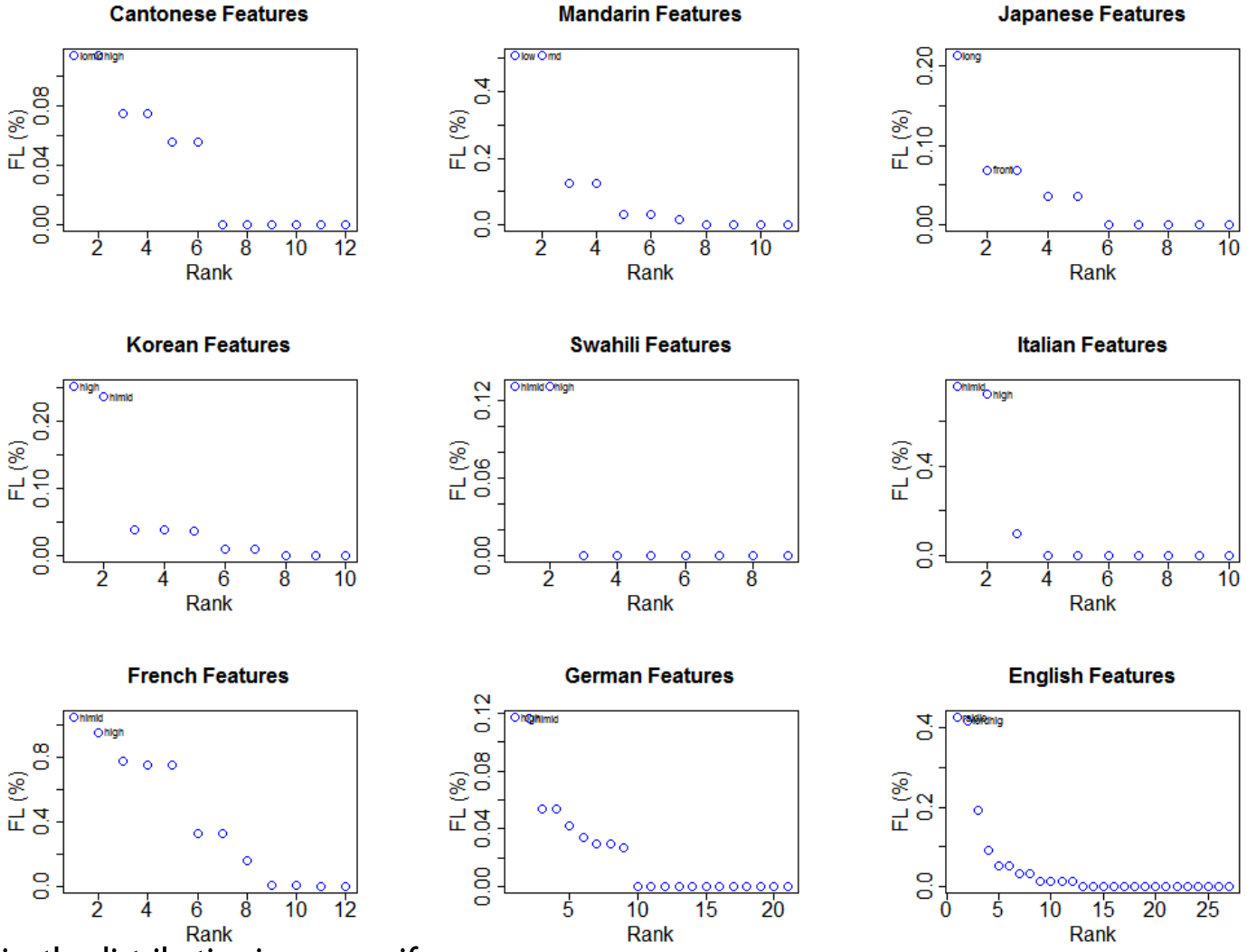
⌦ (Mostly articulatory) description of segments in terms of features based on UPSID (Maddieson & Precoda, 1990)

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

ASLAN UNIVERSITE DE LYON

DDL

FL OF VOCALIC FEATURES ACCORDING TO RANK



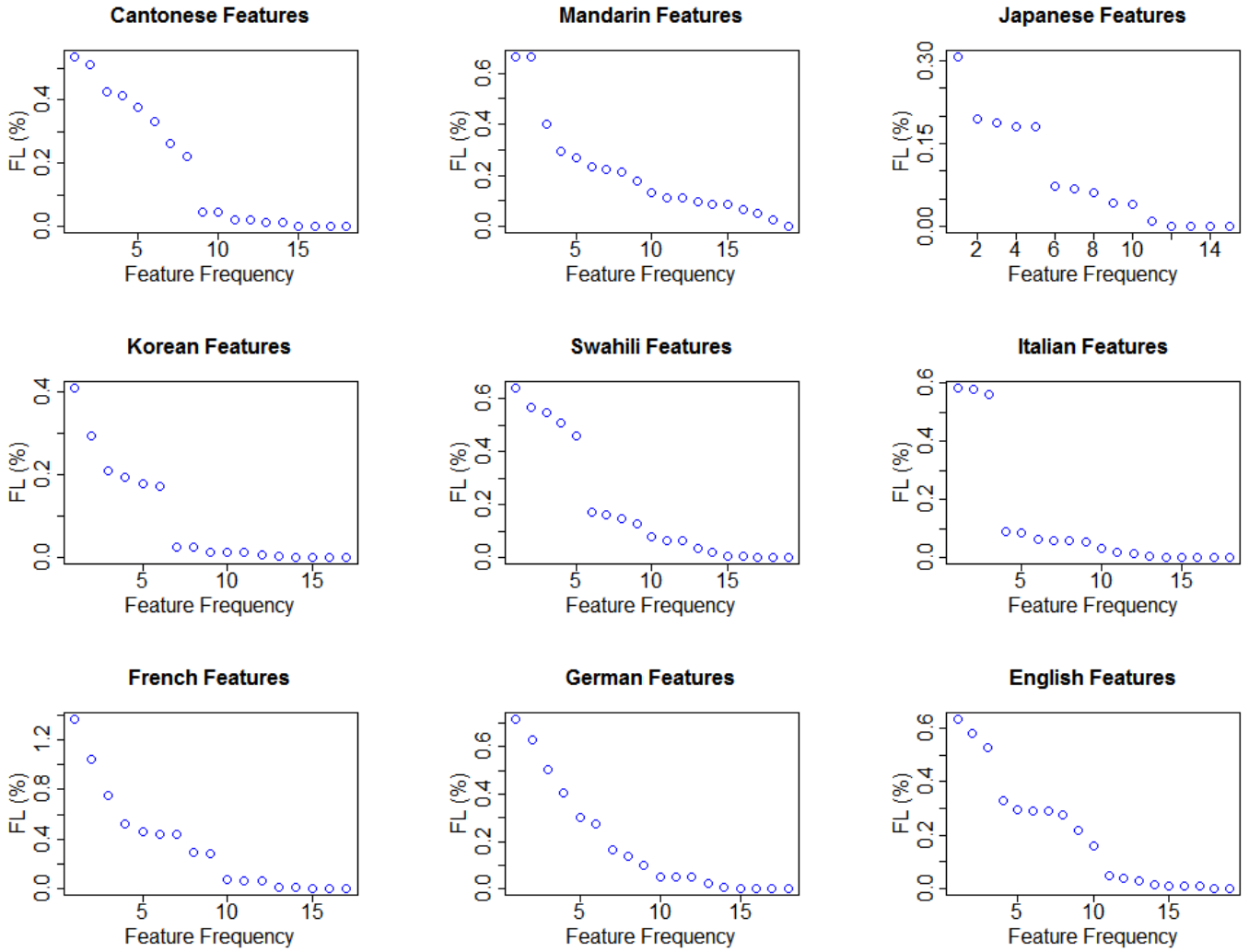
⌦ Again, the distribution is never uniform.

⌦ Often, 1 or 2 features have higher FLs (consistent combinations in terms of dimensions)

ASLAN UNIVERSITE DE LYON

DDL

FL OF CONSONANT FEATURES ACCORDING TO RANK



ASLAN UNIVERSITE DE LYON

DDL

Discussion

Cross-linguistic trends

- ✓ Whatever the organizational scale, phonetic units do not evenly carry the same FL:
  Few units carry a heavy load, while most of the others only carry a very light load
- ✓ Boundary between heavy vs. light load units may be more (for Vs) or less (for Cs) categorical
- ✓ On average, ~50% of the lexical distinctions relied on infra-syllabic components (C, V, T)

➔ In the lexicon, balance between
localized short-term information (measured by infra-syllabic FL) and longer term information

Cross-linguistic diversity

- ✓ Languages differ in their heavy-loaded units [not shown here]
- ✓ Relationship between rank and FL may be linear or logarithmic

Perspectives

- ✓ Relationship between usage frequency and functional load
- ✓ Existence of cross-language trends in the favored heavy-units
  - Importance of coronal consonants and low vowels [not shown here]

ASLAN UNIVERSITE DE LYON DDL

Conclusion

Functional Load sheds some new light on phonological systems

- Robust and multiscale trend towards uneven FL distribution within phonological systems
- Descriptive symmetries may not be reflected as functional symmetries (see 5-vowel systems comparison)
- Identification of well-known motivations (perceptual distinctiveness, ease of articulation) not straightforward…
- Towards a functional account of (phonological) complexity

ASLAN UNIVERSITE DE LYON DDL

Speculative Conclusion

Strong tendency toward an uneven distribution of FL

- Not optimal in a strict information-theoretical framework
  - ✓ But compatible with a high level of redundancy
- Uneven distributions often found in languages (Zipf law, etc.)
  - ✓ (Maximum) re-use of phonological 'chunks' ?
  - ✓ Structure of the lexicon and morphology (small-world network, preferential binding, etc.)
    - ✓ *Distinctiveness vs. efficiency (Kello & Beltz, 2009)*
    - ✓ *Notion of kernel word network (Ferrer i Cancho & Solé, 2001; Dorogovtsev & Mendes, 2001)*
- Existence of a kernel phonemic network?
  - ✓ Heavy-load phonemes and contrasts vs. others
- In this view the latter are not useless because they probably reflect the adaptive nature of the language

Back to Research questions

How is (lexical) information distributed in phonological systems?

Unevenly,
intensively tapping a few contrasts and under-using others

Could it change our view on phonological systems in a typological perspective ?
It's up to you!

Does it tell us something interesting on languages (and speakers) ?

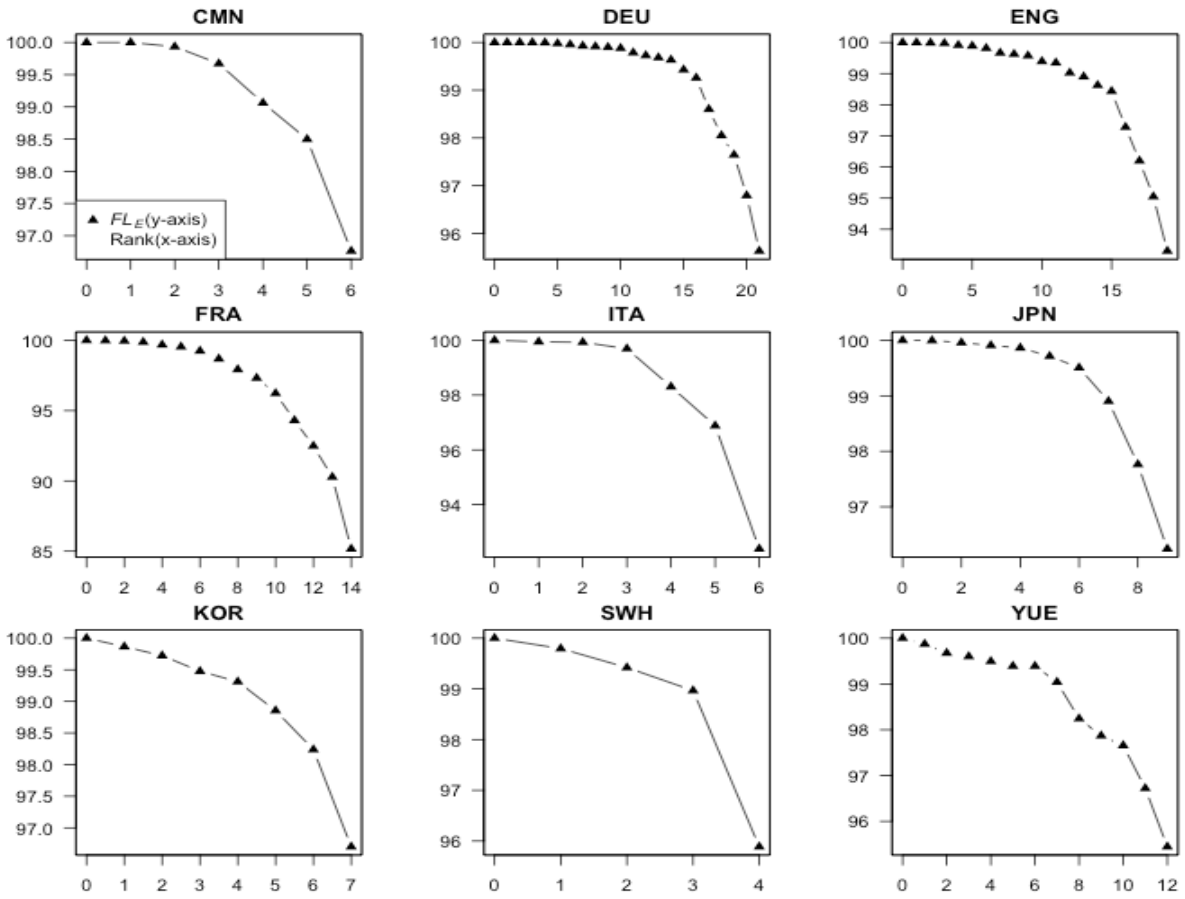
- ➔ *May be useful to address diachronic issues*
- ➔ *May partly explain some variability in psycholinguistics*
- ➔ *Shed light on the Consonantal Bias hypothesis*

WHAT IF…

WE GET RID OF THE LIGHT-LOAD SEGMENTS?

Simulation of the Relative Loss of information Induced by Reducing Vowel System

Remaining vocalic information

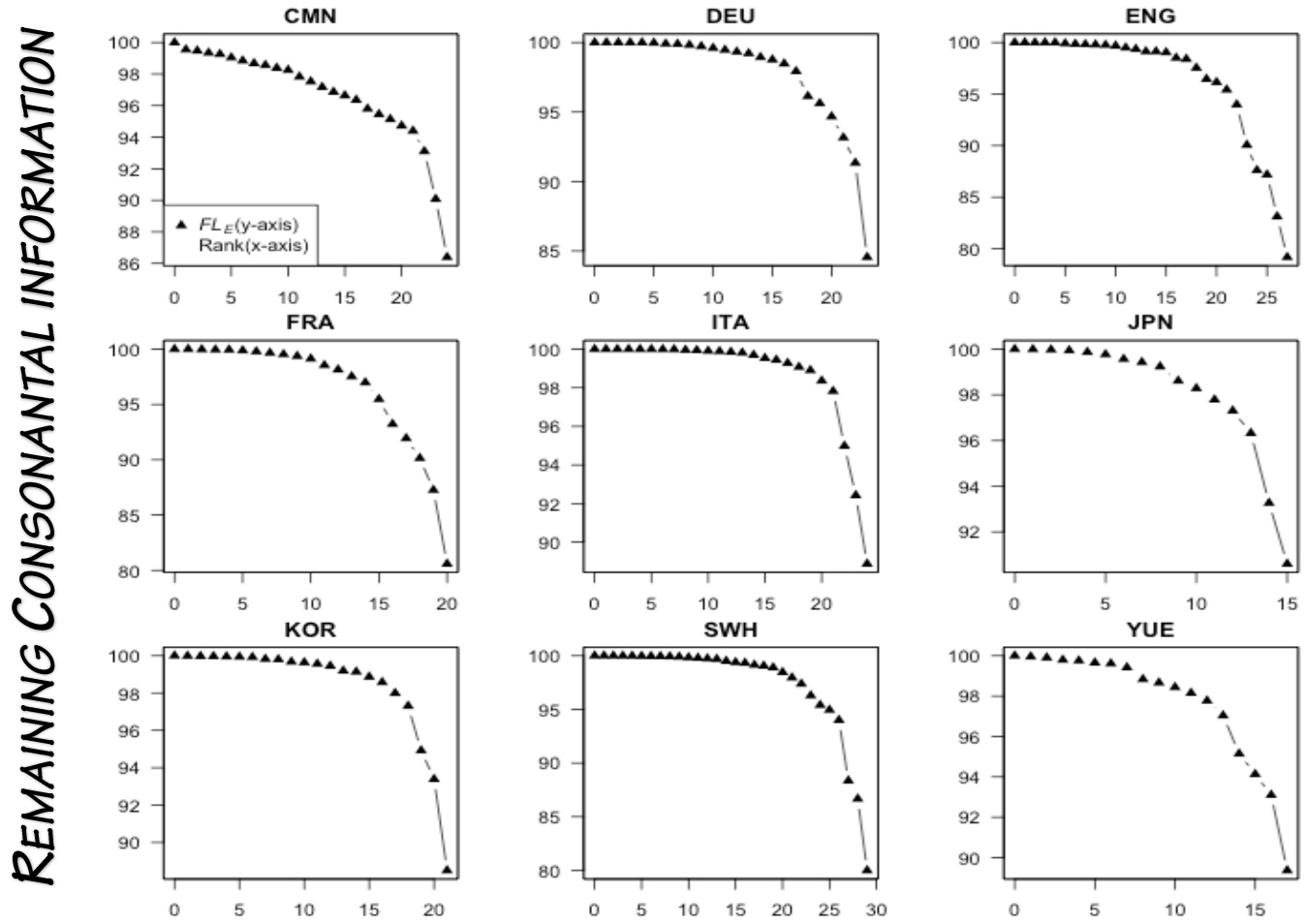


% of FL on the y-axis, vowels listed by their increasing order of FL (x-axis)

ASLAN UNIVERSITE DE LYON

DDL

Simulation of the Relative Loss of information Induced by Reducing Consonant System



% of FL on the y-axis, consonants listed by their increasing order of FL (x-axis)

ASLAN UNIVERSITE DE LYON

DDL

THANK YOU!

və 'nɔv 'wɪnt ən və 'sən wə tɪs'pwətɪn 'wɪtʃ wəz və 'strɔnkə
wæn ə 'tlævlə keɪm ə'lɔn 'læbt ɪn ə 'wɔm 'klɔk
veɪ ə'krit vət və 'wən ɔ 'vəst sək'sɪtɪt | ɪn 'meɪkɪn və 'tlævlə teɪk ɪz 'klɔk ɔv
ʃɔt bɪ kən'sɪtət 'stlɔnkə vən vɪ 'əvə
væn və 'nɔv 'wɪnt 'blɔ əz 'ɔt əz ɪ 'kɔt | bət və 'mɔ ɪ 'blɔ
və 'mɔ 'klɔslɪ tɪt və 'tlævlə 'vɔlt ɪz 'klɔk ə'lɔnt ɪm
ənd ət 'lɔst və 'nɔv 'wɪnd 'keɪv 'əp vɪ ə'tæmpt
væn və 'sən 'ʃɔn 'æt 'wɔmlɪ | ənt ɪ'mɪtɪətlɪ və 'tlævlə 'tɔk 'ɔv ɪz 'klɔk
ənt sɔ və 'nɔv wɪnt | wəz ə'bleɪdʃd tə kən'dæs | vət və 'sən wəz və 'stlɔnkə əv və 'tɔ

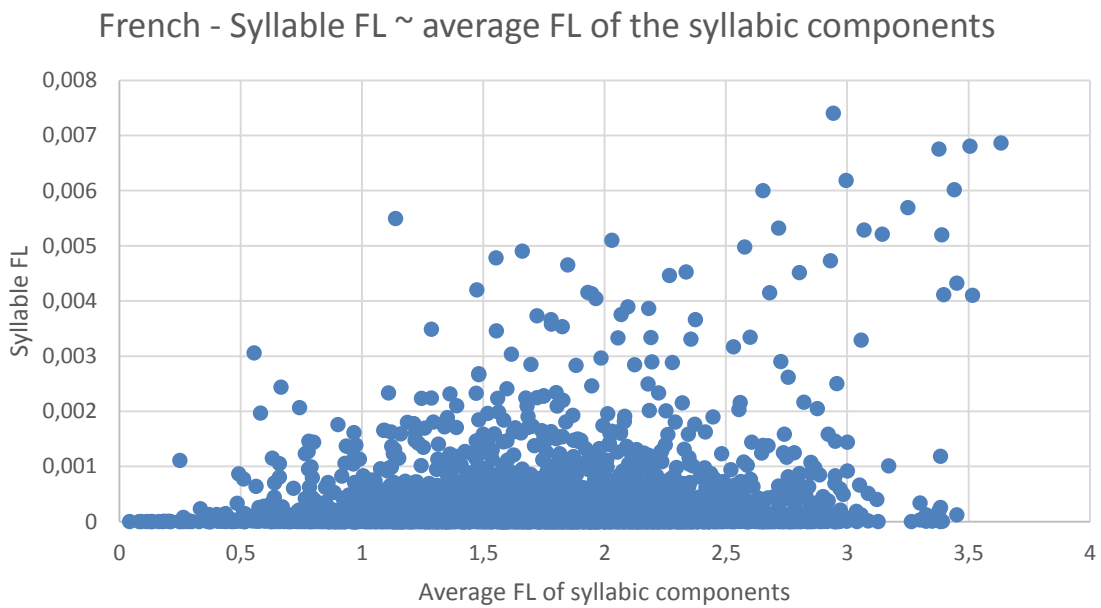
Hint: British English reduced to 11 consonants, 4 vowels, and 1 diphthong

THE NORTH WIND AND THE SUN

ðə 'nɔθ 'wɪnd ən ðə 'sʌn wə dɪs'pjutɪŋ 'wɪtʃ wəz ðə 'strɒŋgə
wɛn ə 'trævlə keɪm ə'lɒŋ 'ræpt ɪn ə 'wɔm 'kloʊk
ðeɪ ə'grid ðət ðə 'wʌn hʊ 'fɜst sək'sidɪd | ɪn 'meɪkɪŋ ðə 'trævlə teɪk ɪz 'kloʊk ɒf
ʃʊd bi kən'sɪdəd 'strɒŋgə ðən ðɪ 'ʌðə
ðɛn ðə 'nɔθ 'wɪnd 'blu əz 'hɑd əz hɪ 'kʊd | bət ðə 'mɔ hɪ 'blu
ðə 'mɔ 'klouslɪ dɪd ðə 'trævlə 'foʊld ɪz 'kloʊk ə'raʊnd ɪm
ənd ət 'lɑst ðə 'nɔθ 'wɪnd 'geɪv 'ʌp ðɪ ə'tɛmpt
ðɛn ðə 'sʌn 'ʃɒn 'aʊt 'wɔmlɪ | ənd ɪ'midɪətlɪ ðə 'trævlə 'tʊk 'ɒf ɪz 'kloʊk
ənd soʊ ðə 'nɔθ wɪnd | wəz ə'blaɪdʒd tə kən'fɛs | ðət ðə 'sʌn wəz ðə 'strɒŋgə r əv ðə 'tu

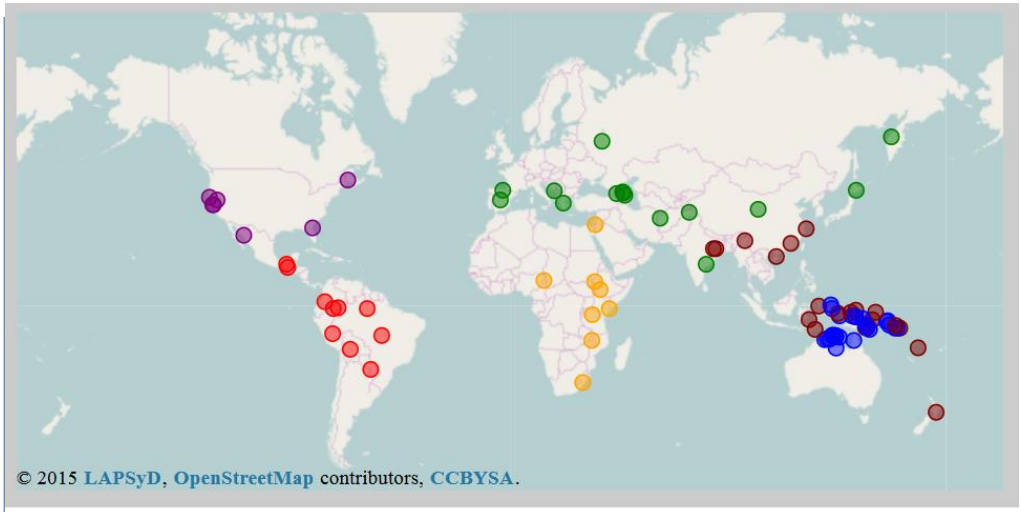
və 'nɔv 'wɪnt ən və 'sən wə tɪs'pwətɪn 'wɪtʃ wəz və 'strɔnkə
wæn ə 'tlævlə keɪm ə'lɔn 'læbt ɪn ə 'wɔm 'klɔk
veɪ ə'krit vət və 'wən ɔ 'vəst sək'sɪtɪt | ɪn 'meɪkɪn və 'tlævlə teɪk ɪz 'klɔk ɔv
ʃɔt bɪ kən'sɪtət 'stlɔnkə vən vɪ 'əvə
væn və 'nɔv 'wɪnt 'blɔ əz 'ɔt əz ɪ 'kət | bət və 'mɔ ɪ 'blɔ
və 'mɔ 'klɔslɪ tɪt və 'tlævlə 'vɔlt ɪz 'klɔk ə'lɔnt ɪm
ənd ət 'lɔst və 'nɔv 'wɪnd 'keɪv 'əp vɪ ə'tæmpt
væn və 'sən 'ʃɔn 'æt 'wɔmlɪ | ənt ɪ'mɪtɪətlɪ və 'tlævlə 'tɔk 'ɔv ɪz 'klɔk
ənt sɔ və 'nɔv wɪnt | wəz ə'bleɪdʃd tə kən'dæs | vət və 'sən wəz və 'stlɔnkə əv və 'tɔ

Multi-scale: Segments & Syllables



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?

5-VOWEL SYSTEMS IN LAPSYD



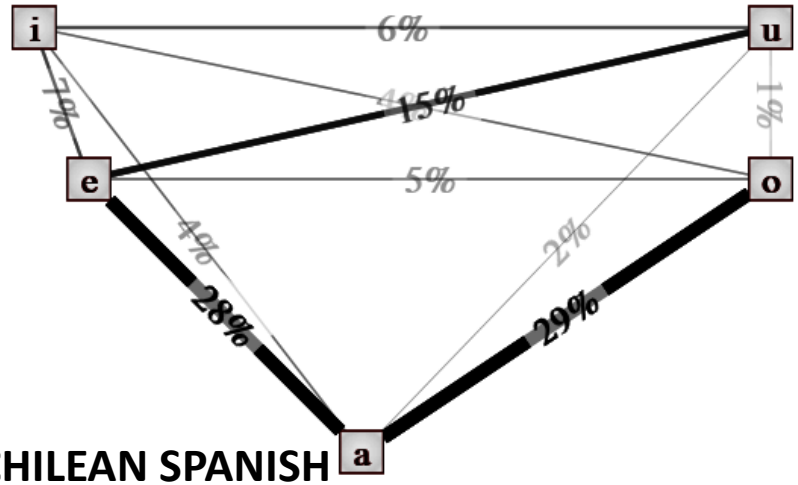
-Europe, W&S Asia
-East and SE Asia
-Africa
-North America
-South America
-Oceania

- 87/633 languages (~13.5%)
  - ✓ Is /i u a e o/ an analysis artefact?
  - ✓ Are these systems really similar?

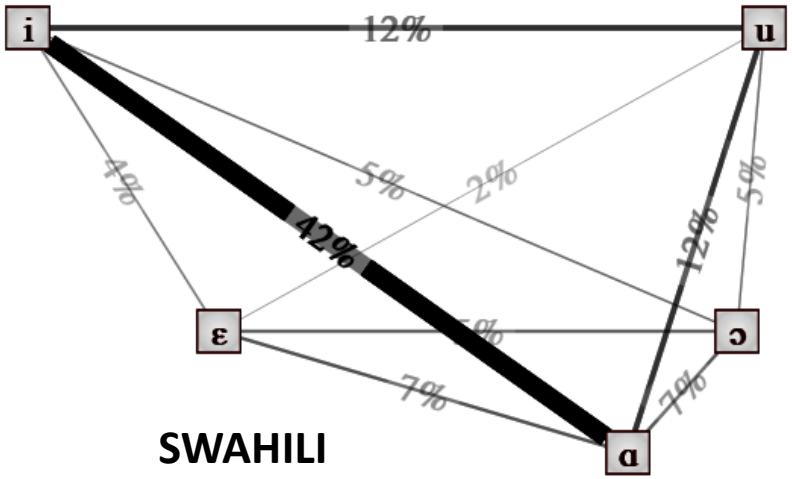
Vowels	front			central		back		
	unrounded		rounded	unrounded		unrounded	rounded	
		long			long			long
high	i [80]	iː [1]		ɨ [5]		ɯ [3]	u [74]	
lowered high	ɪ [2]	ɪː [1]					ʊ [8]	ʊː [1]
higher mid	e [34]	eː [1]	ø [1]	ə [3]			o [33]	
un-mid	'e' [19]			'ə' [1]			'o' [20]	
lower mid	ɛ [28]						ɔ [23]	
raised low	æ [4]			ɐ [2]				
low	a̟ [2]			a [79]	aː [3]	ɑ [2]		

http://www.lapsyd.ddl.ish-lyon.cnrs.fr/

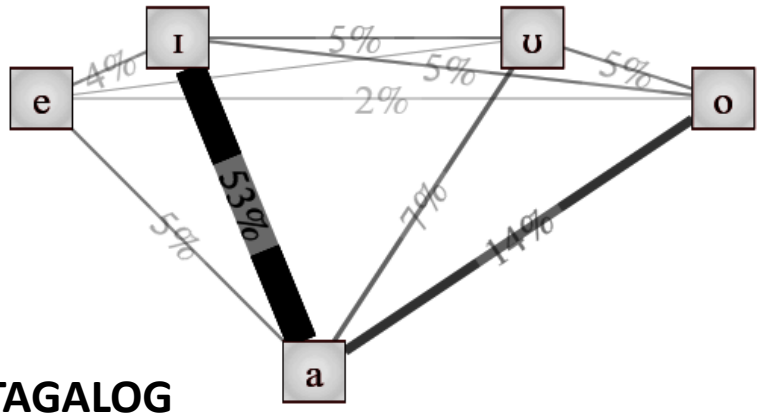
Comparison of 5-vowel systems



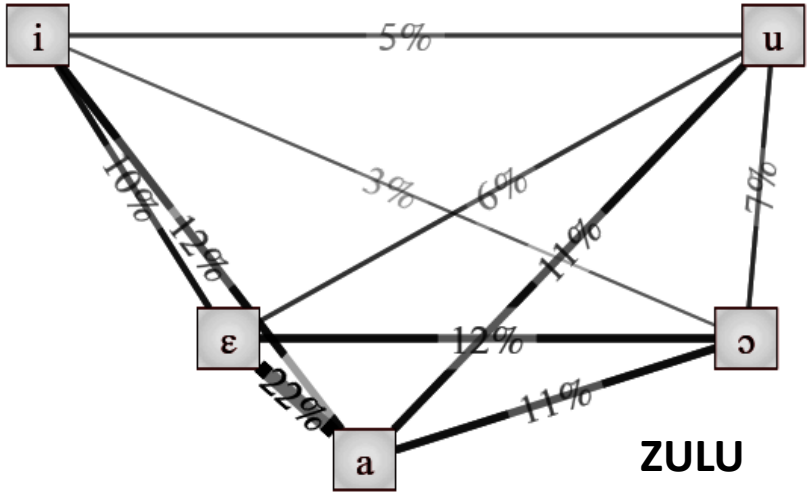
CHILEAN SPANISH



SWAHILI



TAGALOG



ZULU

Highest FL Vowels

	Languages																	
	yue		eng		fra		deu		ita		jpn		kor		cmn		swh	
1	ɔ:	0.71	eɪ	1.12	e	3.63	a	0.71	a	2.34	a	0.76	i	0.58	u	1.73	a	1.02
2	a:	0.66	aɪ	1.00	a	3.51	i:	0.68	e	2.14	e	0.50	a	0.48	i	1.71	i	0.95
3	ɐ	0.65	i:	0.99	ø	2.74	aɪ	0.57	i	1.87	o	0.48	o	0.48	ə	1.66	u	0.45
4	i:	0.45	ɪ	0.93	ã	2.72	ɪ	0.52	o	1.34	i	0.33	e	0.36	a	1.54	o	0.29
5	ɛ:	0.39	æ	0.75	ɛ	2.36	ɛ	0.46	ɔ	0.29	o:	0.25	ʌ	0.27	y	0.54	e	0.24

The 5 vowels largely depart from a canonical system evenly distributed in the vocalic space
The low vowel (/a/-like) is not always the preferred attractor but is present for each language
/i/ or /i:/ are present in 8 out of 9 languages, /e/ and /o/ or /o:/ in 5
/u/ is only present in 2 languages (Mandarin and Swahili)

ASLAN UNIVERSITE DE LYON

DDL

HIGHEST FL CONSONANTS

	Languages																	
R	yue		eng		fra		deu		ita		jpn		kor		cmn		swh	
1	ts	1.36	t	1.74	s	3.40	n	1.49	d	1.07	k	1.26	n	0.79	t	3.44	n	2.18
2	k	1.28	n	1.57	l	3.25	R,r	1.17	l	0.96	s	0.86	g	0.61	l	2.86	j	2.08
3	s	1.08	m	1.35	d	3.14	m	1.03	n	0.81	t	0.79	l	0.51	ʂ	2.85	w	2.01
4	h	0.96	ð	1.28	m	2.01	d	0.85	s	0.76	n	0.74	s^h	0.46	tʂ	2.53	l	1.35
5	t	0.95	s	1.24	n	1.93	z	0.74	k	0.46	m	0.58	d	0.42	p	2.12	z	1.31

Significant presence of coronal consonants with various manners in the first row in all languages (except JPN(/k/))
Relative preference for voiced consonants (27 consonants over all 45 consonants, but 5 over 9 first-rank consonants are voiceless)

ASLAN
UNIVERSITE DE LYON
DDL

From features to articulatory dimensions

i ɯ u

e o

ɛ ʌ

a

Vocalic inventory of Korean

Aperture: 3 sets of mergers

ɛ, e, i
o, u
ʌ, ɯ

Anteriority: 2 set of mergers

ɛ, ʌ
i, ɯ

E.g. To compute the FL of **aperture**, the actual lexicon is contrasted with a lexicon where 3 simultaneous sets of mergers create homophony and modify the distribution of word frequencies.

ASLAN UNIVERSITE DE LYON

DDL

Feature FL: a closer look at dimensions

	V					C			
	Aperture	*Anteriority*	*Roundedness*	*Length*	*Nasalization*	*Place*	*Manner*	*Voicing*	*Aspiration*
yue	0,23	0,15	0,11			1,72	1,07		0,33
cmn	1,02	0,06	0,25			2,07	1,04	0,18	0,67
jap	0,07	0,14		0,21		0,79	0,26	0,36	
kor	0,53	0,08	0,02			0,67	1,01	0,05	0,01
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 also 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.

ASLAN UNIVERSITE DE LYON

DDL

Material: Study #1

Language	Code	# lexical tokens	Corpus coverage	Word distribution Entropy
Amharic	AMH	1.9M	83.7%	12.1
Bulgarian	BUL	6.2M	90.4%	10.5
Chilean Spanish	ChSP	440M	97.7%	9.3
British English	ENG	18M	98.6%	9.5
Estonian	EST	3.4M	84.6%	11.3
Finnish	FIN	970k	72.2%	11.8
French	FRE	900k	98.6%	9.6
German	GER	808k	96.4%	10.1
Swahili	SWA	27.4M	93.6%	10.2
Tagalog	TGL	180k	98.0%	9.5
Turkish	TUR	968k	82.8%	11.8
Zulu	ZUL	217k	75.2%	12.1

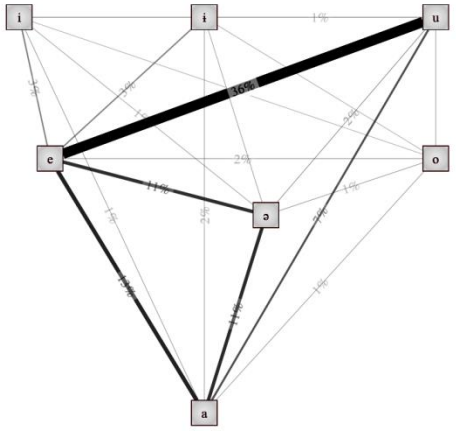
ASLAN UNIVERSITE DE LYON

DDL

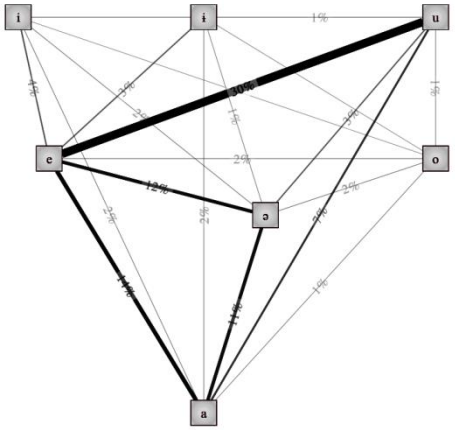
LEXICAL COVERAGE

- Amharic (1.9 M tokens)

20k words (83.7% of coverage)

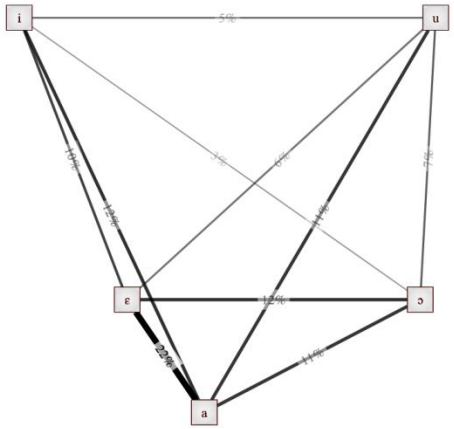


210k words

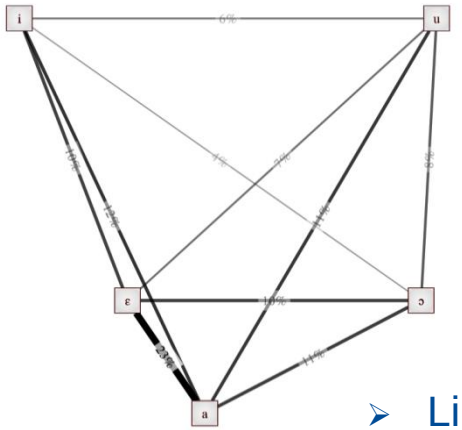


- Zulu (217k tokens)

20k words (75.2% of coverage)



87k words



- Limited impact

TYPE/TOKEN AND INFLECTED/LEMMATIZED WORDFORMS

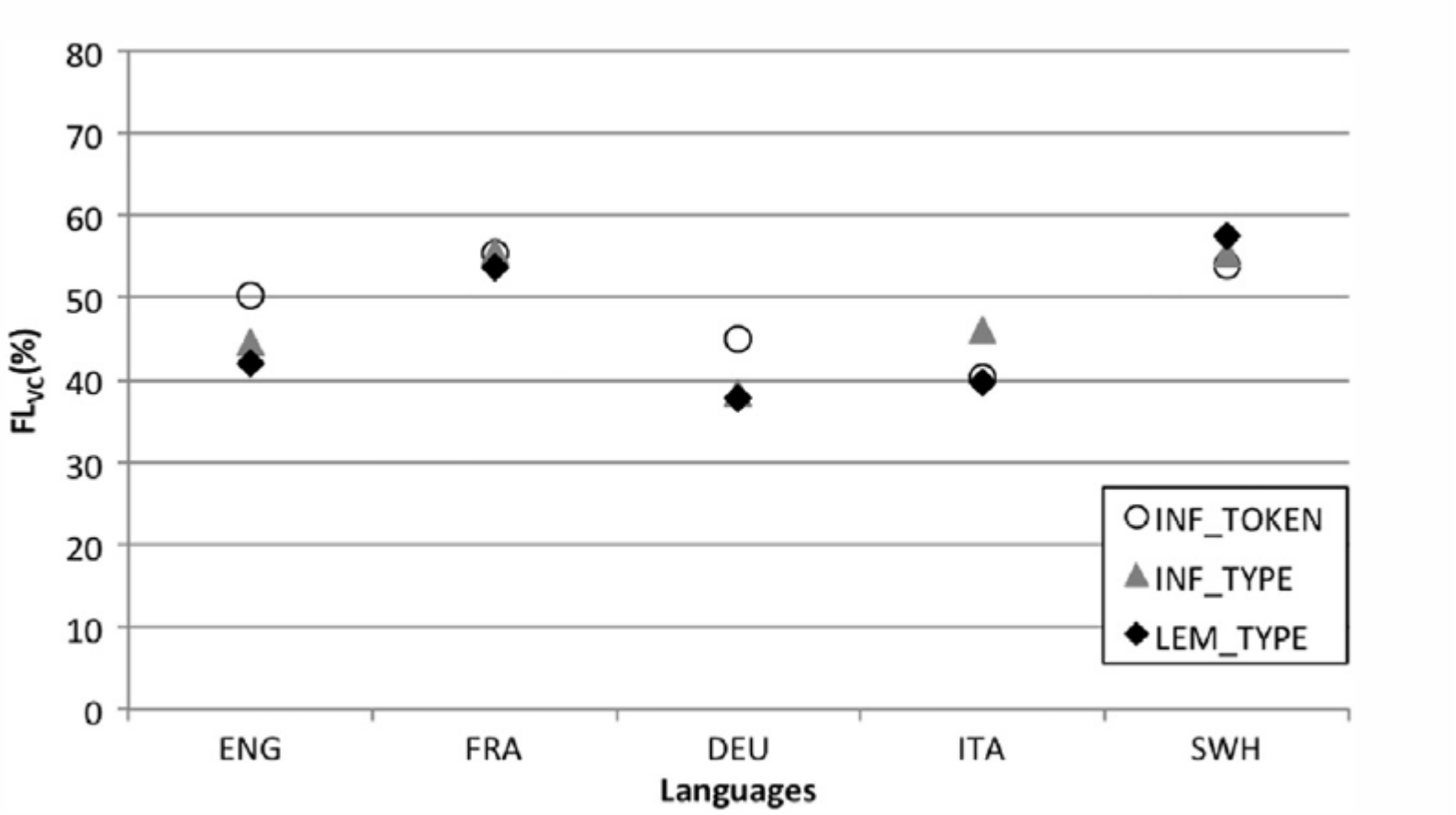


Fig. 3. Segmental functional load (FL_{VC}) in five languages according to corpus configuration.

Oh et al. (2015), *J. PHo.*

Consonantal Bias

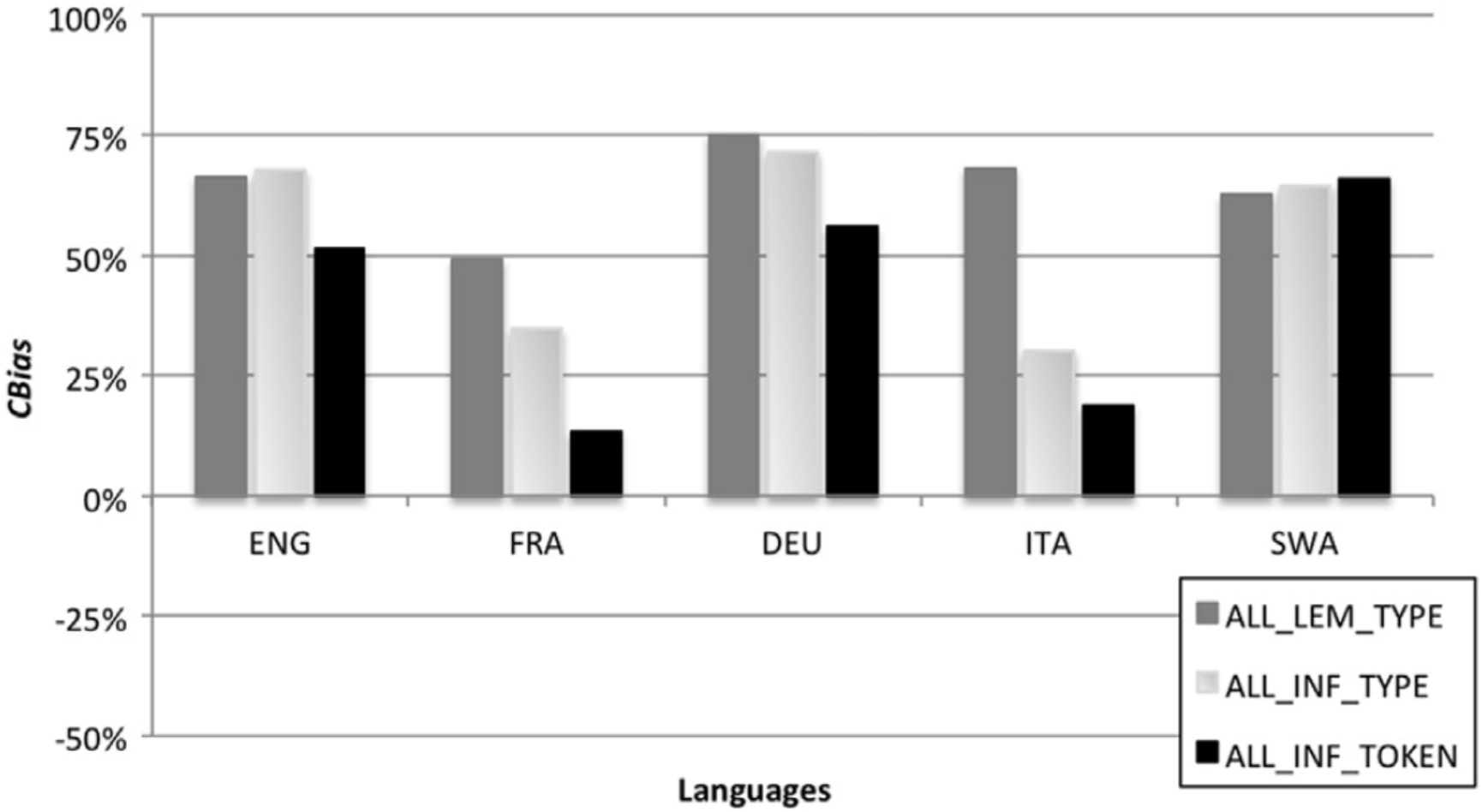
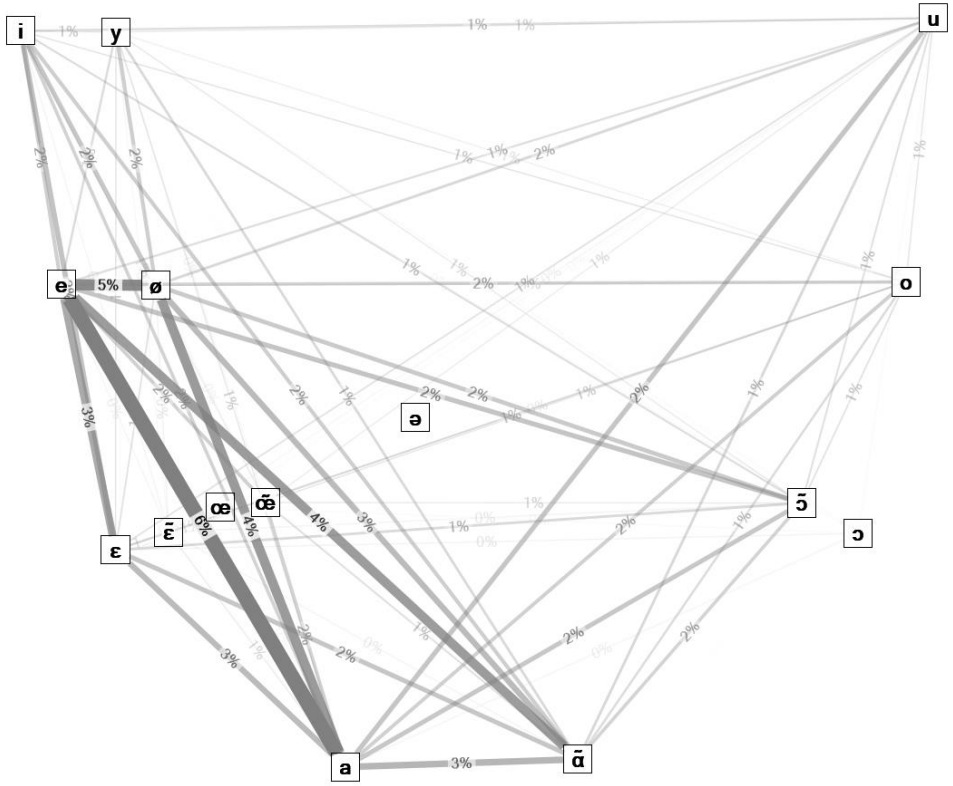


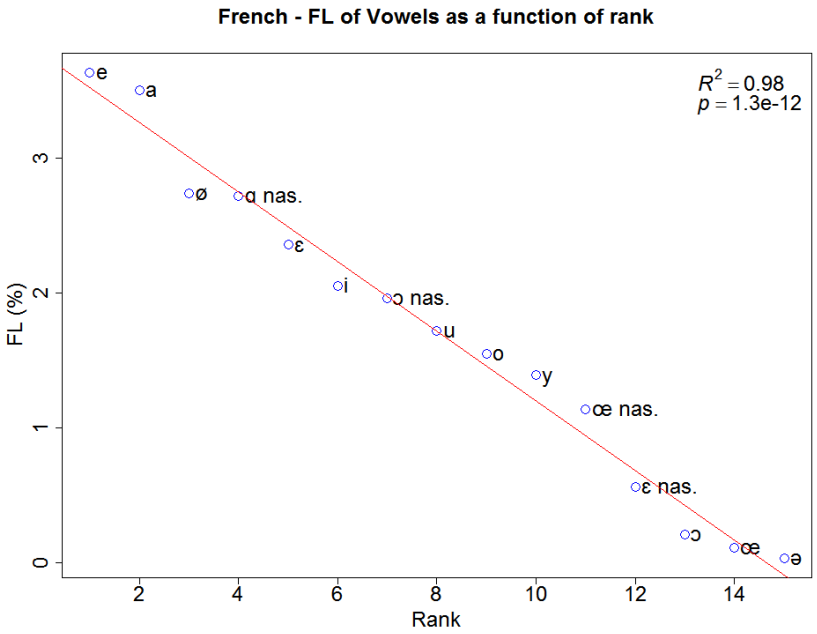
Fig. 4. *CBias* according to corpus configuration.

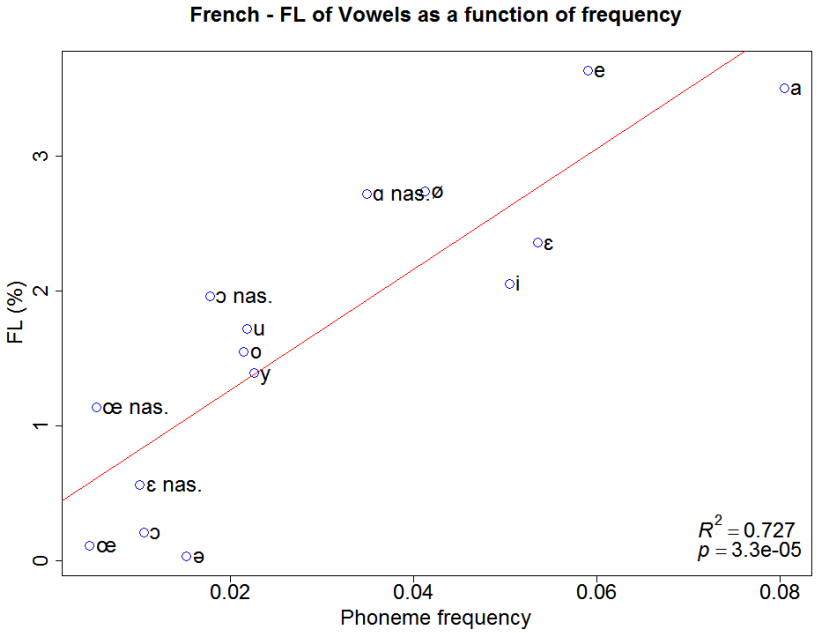
Oh et al. (2015), *J. PHo.*

FUNCTIONAL LOAD AND FREQUENCY

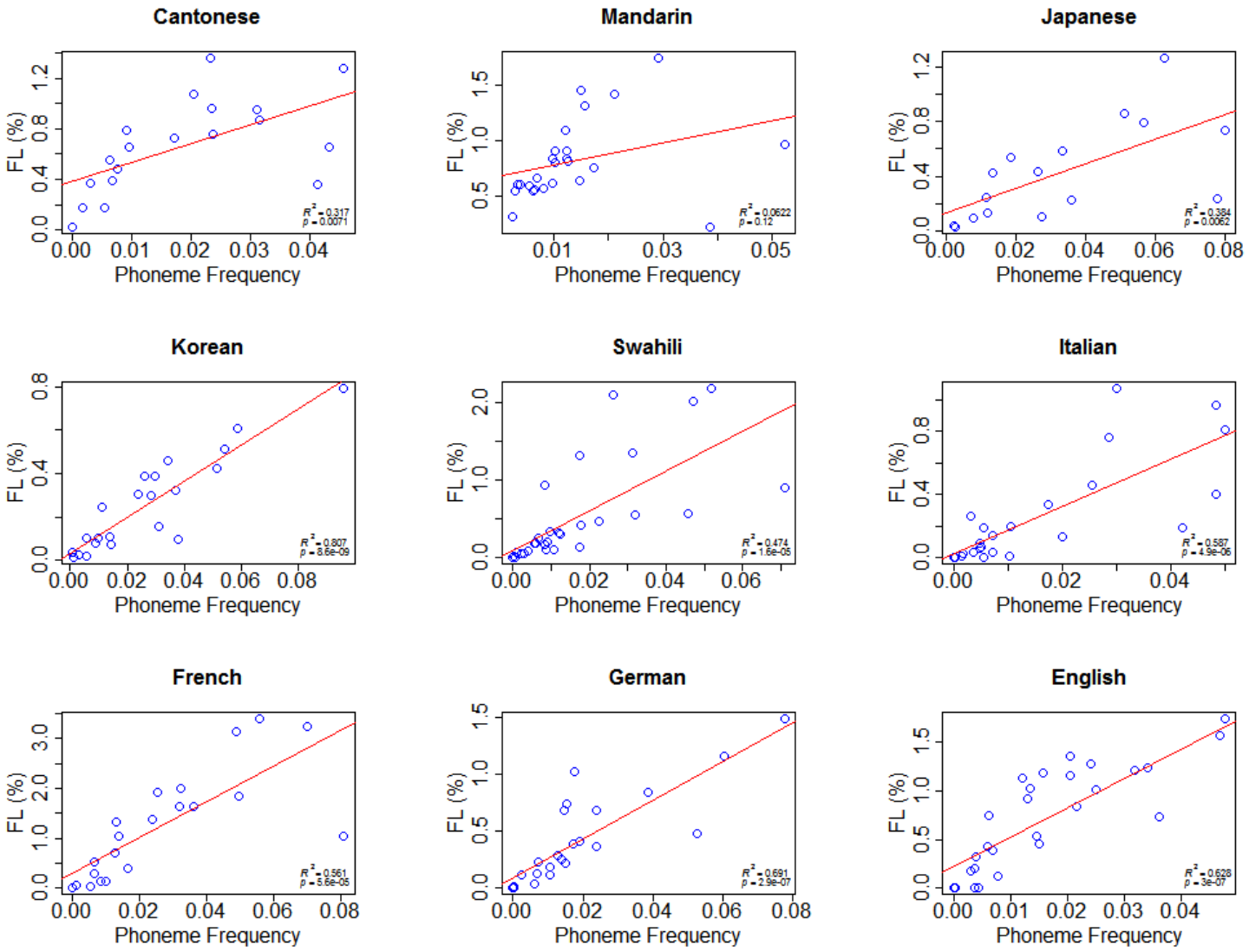
FL of vowels in French







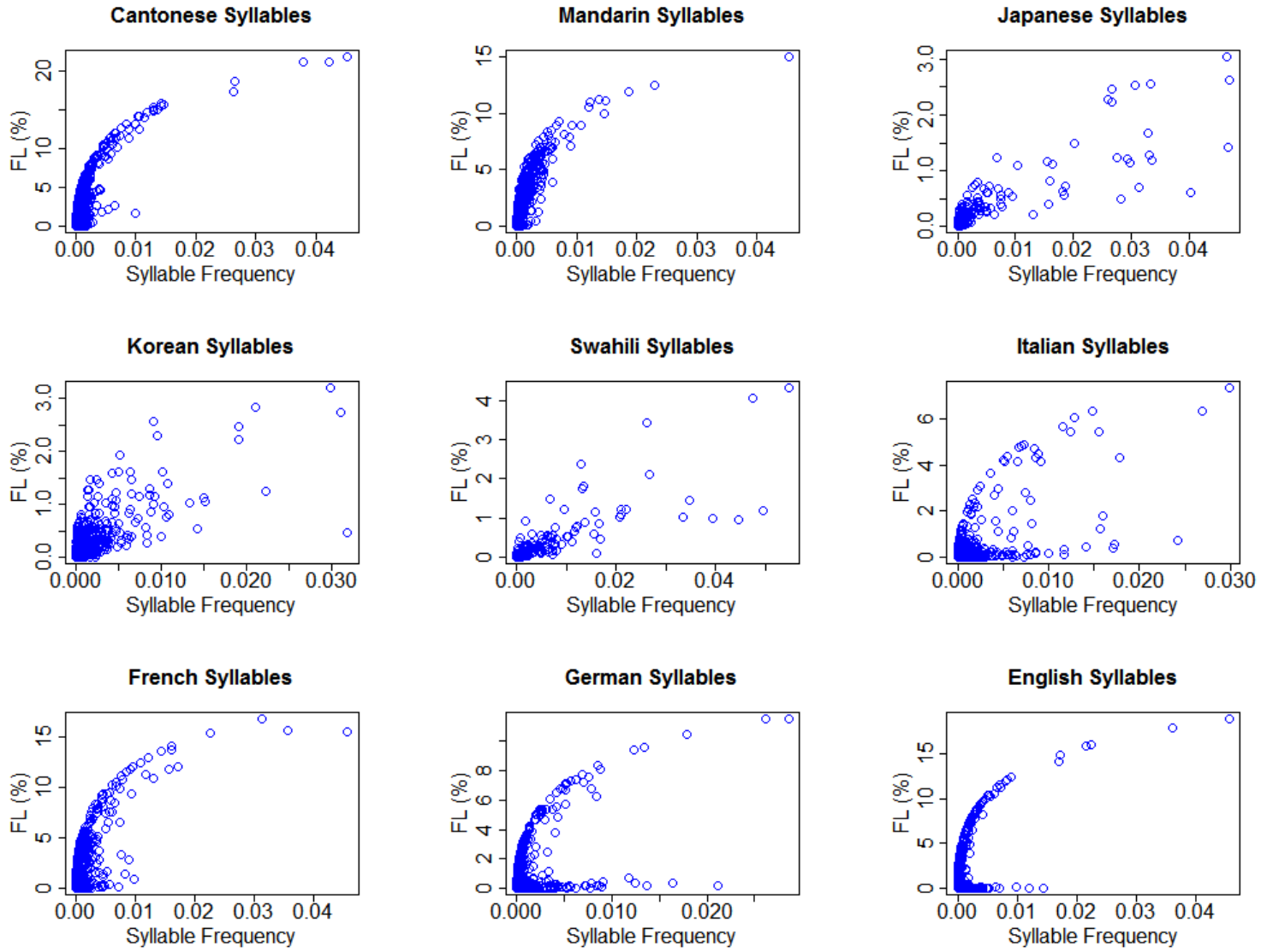
FL OF CONSONANTS AS A FUNCTION OF FREQUENCY



ASLAN
UNIVERSITE DE LYON

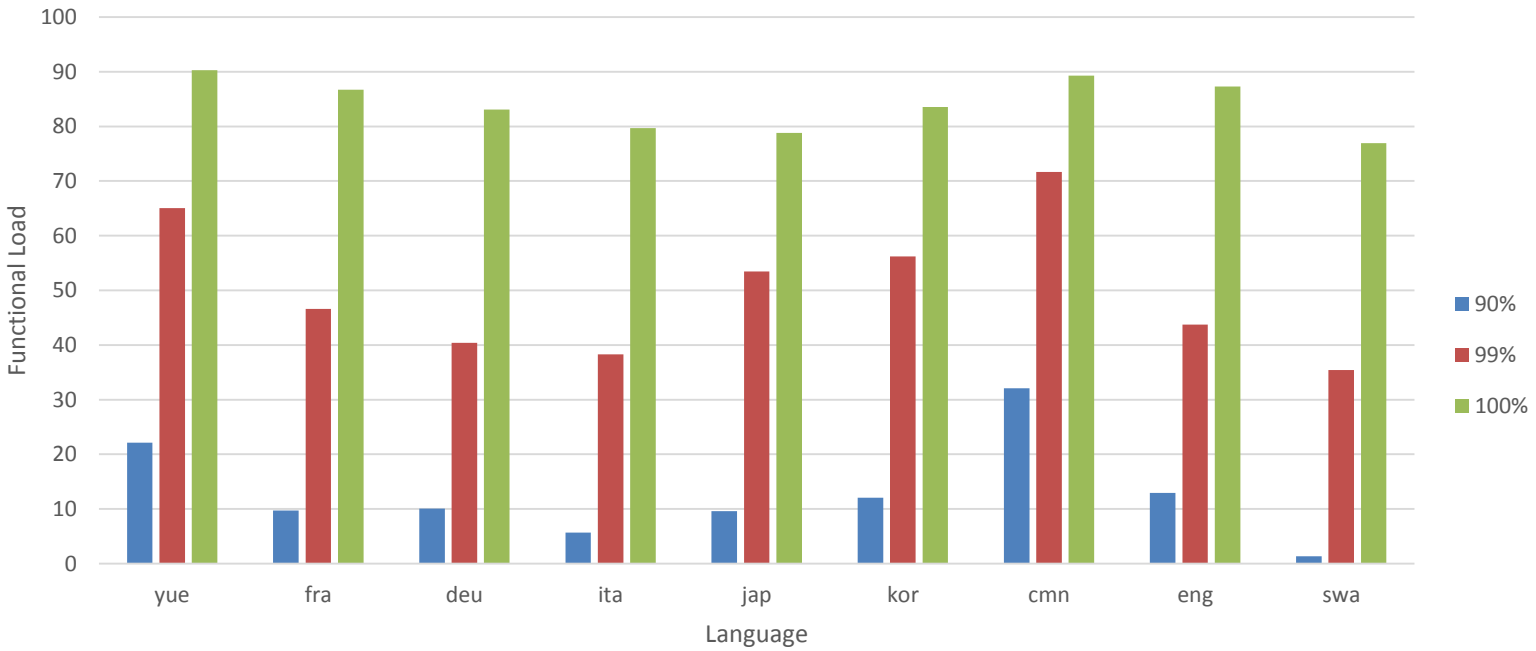
DDL

FL OF SYLLABLES AS A FUNCTION OF FREQUENCY



ASLAN
UNIVERSITE DE LYON
DDL

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

ASLAN UNIVERSITE DE LYON

DDL