This is a draft version. Final version is available here:
Oh, Y. M., et al. "Bridging phonological system and lexicon: Insights from a corpus study of functional load".
Journal of Phonetics (2015), http://dx.doi.org/10.1016/j.wocn.2015.08.003

Bridging Phonological System and Lexicon: Insights from a Corpus Study of Functional Load

Yoon Mi Oh, Christophe Coupé, Egidio Marsico, François Pellegrino

Laboratoire Dynamique du Langage, UMR5596 Université de Lyon and CNRS, France
{yoon-mi.oh; francois.pellegrino}@univ-lyon2.fr,
{christophe.coupe; egidio.marsico}@cnrs.fr

1. Introduction

1.1 The Concept of Functional Load

As stated by Hockett, "The function of a phonemic system is to keep the utterances of a language apart" (Hockett, 1966:1).Phonemes are thus considered the elementary bricks on which contrasts between words are built. The most obvious procedure to identify them is by listing minimal pairs (when they exist): two sound sequences associated with two different meanings and differing by only one element. The set of such 'distinctive' elements constitutes the phonemic system of a particular language. For decades, studying phoneme inventories has been the gateway for understanding how languages work. This traditional approach to phonemes and relations between them has yielded highly significant insights into the organization of phonological systems (Crothers, 1978; Hall, 2011; Hyman, 2008; Liljencrants & Lindblom, 1972; Lindblom, 1986; Lindblom & Maddieson, 1988; Maddieson, 1984; Marsico et al., 2003; Schwartz et al., 1997; Vallée, 1994). However, a side-effect of this paradigm is that, because all phonemes in an inventory are given the same importance, disregarding their frequency and their role in contrasts[1], certain key phenomena remain underappreciated. To illustrate, consider asking a British English (RP: Received Pronunciation) speaker to provide an example of a minimal pair based on a consonantal contrast. Her answer is likely to include word pairs that exhibit a "high frequency" contrast such as /t-d/ (as in "*tip"* vs. "*dip")*, as opposed to word pairs that exhibit a "low frequency" contrast such as /ʒ-v/, (as in "*closure"* /ˈkləʊʒə/ vs. "*clover"* /ˈkləʊvə/). The point is that some phonemic contrasts in English, differentiate hundreds of word pairs (e.g. /t-d/) while others may only be involved in a handful of word pairs (e.g. /ʒ-v/). This fact accords with Hockett's addendum to his characterization of the functional role of phonemes: i.e. that "Some contrasts between the phonemes in a system apparently do more [keeping apart of words] than others" (Hockett, 1966:1). Moreover, this observation appears to hold true for other languages as well, with the work done by particular contrasts potentially varying across languages. Indeed, the Prague School thought that specific contrasts may differ from one language to another and that this "rendement fonctionnel" or "charge fonctionnelle" (Functional Load, henceforth FL) should be taken into consideration when reasoning about phonological systems (Cercle Linguistique de Prague, 1931; Jakobson, 1931).

[1] Even if vowels and consonants (as well as their natural subsets: stops, fricatives, etc.) are not considered identical, in terms of production (Ladefoged & Maddieson, 1996), acoustics (Fogerty & Humes, 2012; Ladefoged, 2001; Stevens, 2002; among others), and perception (Fry et al., 1962; Kronrod, Coppess & Feldman, 2012; Liberman et al., 1957). These differences have recently been mirrored by neurophysiological findings (Caramazza et al., 2000; Mesgarani et al., 2014; Obleser et al., 2010; Scharinger, Idsardi, & Poe, 2011). Vowels and consonants are not identical in terms of functional role either (Nespor, Peña, & Mehler, 2003; New, Araújo, & Nazzi, 2008; Toro et al., 2008), should it be defined by usage frequency or FL, for instance.

1.2 Some Landmarks on Functional Load

Despite a general agreement on what it covers, it should be noted that the concept of FL has often been considered in an impressionistic way (for a review, see Surendran, 2003). As a consequence, FL is generally described by circumlocutions and no precise theoretical definition exists, beyond general statements such as "The term FUNCTIONAL LOAD is customarily used in linguistics to describe the extent and degree of contrast between linguistic units, usually phonemes" (King, 1967). To be fair, one should also note that formal mathematical definitions arose as early as the mid-fifties (Hockett, 1955) and provided enough ground to address FL-related issues. Before this quantitative characterization, advocates of FL heavily relied on intuitions and extensions of the notion of phonological contrast. As stated in the previous section, phonological contrast and opposition were central concepts within the Prague School. Trubetzkoy later mentioned that an "economical" language would very often distinguish words by only one phoneme while "prodigal" languages would make usage of several phonological elements to keep words distinct (Trubetzkoy, 1939:240). Kučera (1963) compared phonemic and syllabic inventory entropies, as well as some derived FL measures, in Russian and Czech. Yet, references to FL have remained sporadic for decades, probably because of the difficulty to process large corpora, which were moreover hardly available. This state lasted until Surendran and Niyogi breathed new life into the concept at the beginning of this century. They compared FL of tones, stress, phonemes and phonetic features in four languages (Dutch, English, German, and Mandarin) and highlighted the importance of the tonal system in Mandarin (Surendran & Niyogi, 2003). This result was confirmed in a follow-up study (Surendran & Levow, 2004) and recently extended to Cantonese (Oh et al., 2013). Oh and colleagues also compared the relative functional weight of consonantal, vocalic (and tonal, if any) systems in five languages (Cantonese, English, Japanese, Korean, and Mandarin). Their results suggest that the distributions of FL in a phonological system are very uneven, with only a few prominent contrasts. These differences in relative prominence may be useful to take into consideration for foreign language acquisition (following Brown, 1988; Munro & Derwing, 2006).

Besides typology-oriented studies, the main topic for which FL was considered relevant was historical linguistics. Upon its inception, Martinet promoted the notion of FL, suggesting that it may play a role in language change (Martinet, 1938; 1955). According to his hypothesis, also adpoted later by Hockett (1966), phonemes involved in high-FL contrasts would be less prone to merging than those involved in low-FL contrasts. Corpus-based studies have failed to confirm this hypothesis for decades (King, 1967; Surendran & Niyogi, 2003; Surendran & Niyogi, 2006), but a recent cross-language study brought some support to it (Wedel, Kaplan & Jackson, 2013). Such conflicting results may be due to differences in corpora or to the small number of sound changes considered so far. It is also possible that, even if FL plays a role in phonetic change, its magnitude is limited, for example with regard to social factors (Labov, 2001). As a consequence, even if FL does determine a pool of potential changes, their actual implementation in a language or a dialect probably depends on further aspects.

From a different angle, the availability of corpora in the field of child language acquisition also stimulated interest in the notion of FL. Its impact on the order of phoneme acquisition by children was demonstrated (Pye, Ingram, & List, 1987; Van Severen et al., 2012), in conjunction with language-specific properties (Stokes & Surendran, 2005). Again, FL is not the only factor at play in the course of phonological acquisition, but converging cues indicate that the phonemes involved in high-FL oppositions have a tendency to be acquired earlier than the others (Van Severen et al., 2012). Stokes and Surendran (2005) showed nevertheless that the effect of FL should be considered with caution since FL was not a significant predictor of consonant order of acquisition in Cantonese-speaking children, in contrast with what they observed in English-speaking children (Stokes & Surendran, 2005).

This re-emergence of the concept of FL can be seen as part of a general movement for promoting statistical and information-theoretic quantitative approaches (see Goldsmith, 2000). Today for instance, the relevance of usage frequency is well acknowledged, and many studies in psycholinguistics, phonology, and phonetics have proven that it significantly impacts cognitive

processes, such as access to mental representations (Bybee, 2003; Cholin, Levelt, & Schiller, 2006; Jescheniak & Levelt, 1994; Johnson, 1996; Levelt, Roelofs, & Meyer, 1999; Pierrehumbert, 2001; Schilling, Rayner & Chumbley, 1998; Walsh et al., 2010). It has nevertheless been less often mentioned in the study of phonological systems per se. However, we think that taking this functional approach into consideration can notably change our vision of phonological systems and can enrich our knowledge of speech cognitive processing. The goal of this paper is consequently to shed new light on phonological systems from the perspective of FL. The emphasis is placed on both their internal functional organization and their importance in building the lexicon. Results are then discussed on communicative and cognitive grounds, in connection with the main focus of this Special Issue.

For almost one century, FL has thus been suggested as a factor involved in the *acquisition* and the *evolution* of phonological units and systems as well as a *systemic* property rooted in lexical strategies. These three dimensions have in common the fact that they deal with the dynamics of structural and functional relationships among the phonological units which define a phonological system. FL especially provides an additional approach to investigate the nature and dynamics of phonological units in the context of their systemic relations (*cross-references in this Special issue to be added*). The COSMO model introduced by Moulin-Frier et al. (this issue) provides a unifying framework able to address the nature of the cognitive architecture of communicating agents, in light of such systemic relations. From an epistemological viewpoint, Moulin-Frier and his colleagues advocate the implementation of alternative theories of speech communication in COSMO multi-agent simulations, and their testing against properties observed in real phonological systems. In their paper, this procedure is applied to regularities observed in phonological inventories (vowel and consonantal systems) and syllable inventories through multi-agent deictic games. They also mention that their work can be extended to address compositionality, thus requiring more elaborate stimuli for their communicating agents. We consider that FL may bring a new set of cross-linguistic regularities that would be especially relevant for testing extensions of the COSMO framework to lexically-based simulations. We suggest that the FL properties extracted from artificial corpora yielded by multi-agent naming games or similar settings (Steels & McIntyre, 1998) should be compared to properties observed in real human lexicons, beyond what has already been explored at the segmental level.

1.3 Paper Outline

Section 2 introduces the methodology implemented in this paper. In Sections 3 and 4, two directions are proposed to illustrate the potential of FL studies. In the first study, we investigate the structure of a phonological system as it is revealed by the FL of vowels, consonants, stress and tones as whole subsystems. Morphological information available for five languages (British English, French, German, Italian, and Swahili) further leads to evaluate FL sensitivity to several factors. Considering token or type frequencies, word-forms or lemmas may reveal or confirm trends on the function of specific phonological categories. More precisely, it has been shown, at least in some languages, that consonants and vowels tend to be preferentially involved in lexical access – for consonants – or rhythmic and syntactic information – for vowels (Bonatti et al., 2005; Cutler et al., 2000; Delle Luche et al., 2014; Havy & Nazzi, 2009; Nazzi & New, 2007; Nazzi et al., 2009; Nespor, Peña, & Mehler, 2003; New, Araújo, & Nazzi, 2008; Toro et al., 2008). What has been coined Consonant Bias, potentially reflected by FL, will thus be the main issue at stake. In Section 4, the second study focuses on distributions of FL at the level of segmental units rather than phonological subsets. It thus investigates general trends or specificities regarding the internal functional organization of phonological systems in the world's languages. The quantitative measures of FL yielded by the framework suggest that representation of phonological (sub)systems based on frequency/usage (Figure 1, right) maybe as useful as the more traditional, time-tested representations (like Figure 1, left). Indeed, by directly encoding the different functional roles of vowels in terms of number of contrasts, Figure 1 (right) reveals salient differences among vowels. For instance, the near-close vowels /ɪ/ and /ʊ/ behave very differently: /ɪ/ being frequent and engaged in a lot of lexical oppositions while the opposite is observed for /ʊ/. Moreover it gives a view of the *system* as a set of intricate

oppositions among its constituents, rather than a set of apparently independent segments, as in the left chart.

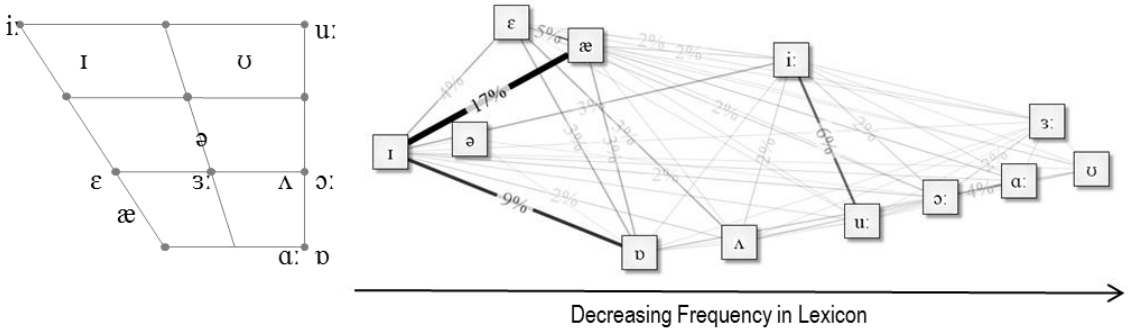

Figure 1. *Illustrations of English (RP) vowel system. Left. Standard IPA chart. Right. Functional network-based representation. Vowels are ranked, from left to right, according to decreasing usage frequency. Edges (thickness and opacity) reflect the functional load associated with each vowel pair. Vertical positions of the vowel labels are arbitrary and chosen for legibility (data computed from WebCelex using the methodology described in Section 2 of the paper).*

Finally, in Section 5, results are discussed in terms of phonological units and features, relative weights of vowel vs. consonant, and general trends in the FL distribution within the phonological systems (see also Oh et al., 2013).

2. Rationale and Methodology

2.1 Computing Functional Load

Several algorithmic approaches have been proposed to quantify FL (Hockett, 1955, 1966; Ingram, 1989; King, 1967; Kučera, 1963; Surendran & Niyogi, 2003; Wang, 1967). Following Hockett (1955), these approaches are grounded in information-theoretic methods (Shannon, 1948) and use entropy computed at various levels as the essential metrics. One noteworthy exception is the simple counting of the number of lexical minimal pairs based on each contrast (Ingram, 1989).

Surendran and Niyogi (2003) and Van Severen and colleagues (2012) thoroughly discussed several of these metrics and the latter showed that Ingram's approach and an entropy-based metric implemented by Surendran and Niyogi (2003) are almost equivalent predictors of the age of acquisition of word-initial consonants in Dutch. However, they differ in the information they encompass and we chose to implement both metrics, referring to them as number of Minimal Pairs (*#MP*) and Entropy FL (FL_E) respectively.

For each language studied, the material consists of a large set of word-forms associated with token frequencies drawn from a large, phonemically-transcribed, corpus. This dataset can optionally be pre-processed in order to filter out specific items (according to their token and lemma frequency, their grammatical category, etc., see Section 2.3). In this paper, the phonological inventory is defined as the pool of phonemes required to transcribe the corpus considered.

For each pair of phonemes in the inventory, *#MP* is the number of distinct word-forms that are discriminated by this specific pair. Because perceptual confusions (in language acquisition) and diachronic mergers (in language change) are more likely to occur between similar phonemes, several studies have limited the inspected contrasts to phoneme pairs that differ only by one phonological feature: place of articulation, manner of articulation or voicing for consonants (Van Severen et al., 2012; Wedel, Kaplan, & Jackson, 2013). However, since our goal was to study the global utilization of the phonological inventory for lexical purposes, no such limitation was implemented and all contrasts were considered. For example, in British English, the lexical items *hit*, *bit*, *pit*, and *sit* contributed to the contrasts /h-b/, /h-p/, /h-s/, /b-p/ /b-s/, and

/p-s/. However, lexical differentiations involving an insertion did not contribute to FL; for instance, the lexical pair *hit-it* did not form a minimal pair.

Besides the Minimal Pair approach, we also implemented the information-theoretic approach proposed by Hockett (1966) and further elaborated by Surendran and Niyogi (2003). Here, a language *L* is considered as a source of sequences made of word-forms *w* taken from a finite set of size N_L and composed of Vowels (possibly including diphthongs), Consonants (possibly including glides) and possibly Stresses and Tones taken from the phonological inventory **P** = **V** ∪ **C** ∪ **S** ∪ **T**. The amount of information of source *L* is estimated in terms of Shannon entropy *H(L)* (Shannon, 1948):

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

where p_{w_i} is the probability of word-form w_i, approximated by its relative token count estimated from the corpus.

Following Surendran and Niyogi (2003), we implemented the definition of FL given by Carter (1987) and derived from Hockett's initial proposal (Hockett, 1966). The FL of a contrast between two phonemes φ and ψ, FL_E(φ,ψ), is defined as the relative difference of entropy between two states of language *L*: the observed state *L* and a fictional state $L^*_{\varphi\psi}$ in which the contrast is neutralized (or coalesced, in Hockett's terminology). FL_E(φ,ψ) therefore quantifies the perturbation induced by merging φ and ψ, in terms of increase of homophony and of changes in the distribution of word frequencies:

$$FL_E(\varphi,\psi) = \frac{H(L) - H(L^*_{\varphi\psi})}{H(L)} \qquad (2)$$

FL(φ, ψ) is hence defined at the level of phonemic *contrasts*, as a ratio theoretically ranging from 0% to 100%.

In addition, one can also focus on the level of *phonemes* themselves, by summing *FL(φ, ψ)* over all the contrasts in which a phoneme φ is involved. *FL(φ)* thus measures the importance of phoneme φ in the language lexical network:

$$FL_E(\varphi) = \frac{1}{2}\sum_{\psi} FL_E(\varphi,\psi) \qquad (3)$$

With the normalization factor ½ applied to ensure that:

$$\sum_{\varphi} FL_E(\varphi) = \sum_{\varphi,\psi\neq\varphi} FL_E(\varphi,\psi) \qquad (4)$$

It can also be used to give a more global quantification of the functional weight of subparts of the phonological system. We defined FL_V (resp. FL_C) as the overall loss of information induced by comparing language *L* with a fictional state L^*_V (resp. L^*_C) in which all vowels (resp. consonants) are merged into a unique symbol. As an illustration, in L^*_V, the three English words *pit*, *bit*, and *pot* coalesce into two forms pVt and bVt while they result in two other forms CɪC and CɒC in L^*_C. Syllable boundaries are taken into account to distinguish between words – e.g. Xī'ān and xīan in Mandarin – and for the computation of FL. For instance, during the computation of FL_C for English, the two words *mattress* /mæ.trɪs/ and *maxim* /mæk.sɪm/ result in two distinct entries /Cæ.CCɪC/ and /CæC.CɪC/, while they would merge into a single entry /CæCCɪC/ if syllable boundaries were not considered.

In addition to FL_V and FL_C, a more drastic reduction was implemented by only keeping the skeleton of the word-forms, i.e. consonantal and vocalic slots as well as stress and syllable boundaries. This so-called segmental FL FL_{VC} measures the cumulative information carried by

the identity of the segments in the wordlist. In the resulting L^*_{VC} language, the three words mentioned above merge into a CVC form.

$$FL_V(L) = \frac{H(L) - H(L^*_V)}{H(L)} \quad (5)$$

$$FL_C(L) = \frac{H(L) - H(L^*_C)}{H(L)} \quad (6)$$

$$FL_{VC}(L) = \frac{H(L) - H(L^*_{VC})}{H(L)} \quad (7)$$

By extension, stresses and tones can also be considered the same way. For instance in Mandarin, the lexical pair 判 ("sentence", /pʰan4/) and 盘 ("plate", /pʰan2/) contributes to the computation of FL_E between tone2 and tone4, and the global functional weight FL_T of the tonal system can thus be quantified mutatis mutandis, and an overall infra-syllabic FL_{VCTS} is also defined. It is important to note that FL_{VCTS} is not the sum of FL_V and FL_C. Although a strict mathematical proof is difficult to formulate, the following explanation can be given. Coalescing at the same time all vowels together and all consonants together necessarily merges all the word-forms that are merged by coalescing vowels only, and all the word-forms that are merged by coalescing consonants only (whether some word-forms merge in both cases is not relevant). Additionally, more mergers may occur between word-forms of similar phonological pattern (eg. CV, CVC, CV CCVC, etc.) that weren't merged either in L^*_C or in L^*_V. Conversely, for FL_{VC} to be equal to FL_V + FL_C, no word-form that did not get merged in either L^*_V or L^*_C should get merged in L^*_{VC}. This imposes strict constraints on the structure of word-forms that natural languages are usually far from respecting. As an example, while the invented language {pi, bi, pa, ba} (with frequencies all equal to 1) satisfy the constraint, the slightly different language {pip, bi, pa, ba} (again, all frequencies equal to 1) does not.

#MP and FL_E differ in several ways, though they yielded similar results in previous studies (Surendran & Niyogi, 2003, Van Severen et al., 2012). For a given contrast φ-ψ, *#MP* only requires a knowledge of the word-forms in which the two phonemes are involved in order to count the relevant minimal pairs. However, *#MP*(φ, ψ) is not influenced by the rest of the lexicon, i.e. word-forms where φ and ψ are absent. It does not rely on any probability estimation either, which leads Wedel et al. to consider it as a *local* measure (Wedel et al., 2012). On the contrary, Entropy FL is a *global* measure. The entropy is computed on the whole lexicon and involves probability estimations. As a consequence, FL_E(φ, ψ) both requires a global knowledge of the lexicon *and* measures the impact of the φ-ψ contrast on the whole lexicon. Beyond the local influences on lexical access (e.g. Luce & Pisoni, 1998), it has been very recently suggested that global properties of the mental lexicon may influence lexical cognitive processing (Vitevitch, Chan & Goldstein, 2014) and further investigations on the relationship between local and global levels will be insightful, though beyond the scope of this paper.

We introduced in this section several indices aimed at assessing the importance of phonological components in the maintenance of lexical distinctions. These components are however complemented with other dimensions: number of segments or syllables, syllabic structures, phonotactic and syllabotactic information, and more generally word structure. In the rest of this paper, we refer to these dimensions as structural information.

2.2 Language Description

Table 1 provides the description of the data and phonological system of the nine languages (Cantonese, English, French, German, Italian, Japanese, Korean, Mandarin, and Swahili) analyzed in this paper. For five languages (English, French, German, Italian, and Swahili), lemmatized forms were available.

The number of vowels (including diphthongs), consonants, tones (if any) and stresses (if any) are provided for each language. The size of the phonological system may not correspond exactly to traditional phonological descriptions since the corpora used here included some loanwords

and newly coined words derived from other languages. For instance, in the Swahili corpus, there are plenty of Arabic and English loanwords which consequently extended syllabic structures beyond traditional "open" syllables (see Appendix 1). Following Maddieson (2013), syllable complexity is estimated by a syllable index, ranging from 1 to 8 among the world's languages. This index corresponds to the sum of the potentially maximal number of onset, nucleus, and coda elements. For this study, indices were retrieved from the LAPSyD website (Maddieson et al., 2013). The four Indo-European languages (English, French, German, and Italian) have complex syllable structures. The two Sino-Tibetan languages, Cantonese and Mandarin, as well as Korean and Japanese, have moderately complex syllable structures. Swahili has simple syllable structures.

Table 1. *Language and Corpus Description. For each language, the size of its phonological system (V: #vowels, incl. diphthongs; C: #consonants; T: #tones; S:#stresses, if applicable), syllable index (based on LAPSyD), and the size of syllable inventory (#distinct syllables) are provided, as well as morphological typology information.*[2]

Language	*ISO 639-3 Code*	*Phonological system*		*Syllable index*	*Size of syllable inventory*	*Morphological type*	*Corpus*
Cantonese	YUE	C V T	19 13 6	3	1 303	Analytic / Isolating	A linguistic corpus of mid-20th c. Hong Kong Cantonese
English	ENG	C V S	25 24 2	8	6 469	Analytic / Fusional	WebCelex
French	FRA	C V	22 15	7	5 530	Synthetic / Fusional	Lexique 3.80
German	DEU	C V S	25 32 1	8	6 867	Synthetic / Fusional	WebCelex
Italian	ITA	C V S	25 8 1	6	1 970	Synthetic / Fusional	The Corpus PAISÀ
Japanese	JPN	C V	16 10	4	484	Synthetic / Agglutinative	The Corpus of Spontaneous Japanese (CSJ)
Korean	KOR	C V	22 8	4	2 319	Synthetic / Agglutinative	Leipzig Corpora Collection (LCC)
Mandarin	CMN	C V T	25 7 5	4	1 378	Analytic / Isolating	Chinese Internet Corpus (S. Sharoff)
Swahili	SWH	C V	30 5	2	1 447	Synthetic / Agglutinative	(Gelas, Besacier, & Pellegrino, 2012)

The small sample considered here also provides some variation in terms of morphological type.

Morphological typology deals with the internal word structures. Languages are usually categorized along two dimensions: i) the internal complexity of words in terms of number of morphemes and ii) the assembling strategy for these morphemes. These two dimensions give rise to several morphological language types (Aikhenvald, 2007).

[2] The phonemic inventories of the nine languages (obtained from each corpus) are given in Appendix 1.

Regarding the number of morphemes per word, linguists distinguish between analytic and synthetic languages[3]. Analytic languages tend to limit the number of morphemes they pack in each word, a one-to-one correspondence being the norm. Synthetic languages on the contrary, make frequent use of words consisting of several morphemes. This distinction should be seen as a continuum, ranging from strictly analytic languages (e.g. Vietnamese) to languages where most words consist of several morphemes (e.g. Korean). Between them, one finds languages that lean towards analytic behavior (e.g. English has a tendency to have a low number of morphemes per word) or towards synthetic word formation (e.g. French and Italian are moderately synthetic).

With regards to the assembling strategy, the strict analytical languages have only one morpheme per word and they are thus said to be isolating. Languages that allow or impose several morphemes per word fall into two categories: Agglutinative languages (such as Korean and Japanese) have a strong tendency to maintain clear boundaries between these morphemes. In agglutinative languages, a word typically consists of a sequence in which each morpheme is clearly identified and carries one semantic feature (e.g. number, case, gender). In fusional languages, on the contrary, several semantic features may be merged into one morpheme and it may be difficult to identify the morphemes from the word-form. Romance and Germanic languages are fusional to some degree.

These categories of word formation only provide an outline that cannot account for the richness of morphological processing, both in terms of verbal vs. nominal domains or derivational vs. inflectional dimensions. For instance, both French and German are classified as synthetic / fusional languages, but nominal morphology is more elaborated in German than in French because of the case-marking system. In the rest of this paper, we only scratched the surface of this richness by comparing the FL patterns obtained with corpora consisting of lemmas vs. inflected forms, in order to shed light on potential differences between lexical and grammatical (bound) morphemes.

2.3 Data and Preprocessing

For each corpus, the first step consisted of discarding erroneous word-forms (including non-alphabetical characters). Then, a specific preprocessing was applied as a function of the corpus nature.

For Mandarin, the Chinese Internet Corpus (Sharoff et al, 2006) was retrieved online. For Cantonese, the *Linguistic corpus of mid-20th century Hong Kong Cantonese* (Research Centre on Linguistics and Language Information Sciences, 2013) was also downloaded. For both languages, public domain dictionaries and software - the CC-CEDICT dictionary (CC-CEDICT, 2012) and NJStar Chinese Word Processor (NJStar Software Corp, 2013) for Mandarin and CantoDict (Sheik, 2013) and JyutDict (Learner, 2013) for Cantonese - were used to get the pinyin and jyutping transcriptions respectively. For Mandarin, the transcription software was used when an entry of the corpus was missing in the dictionary. For Cantonese, the transcriptions provided by the two dictionaries were compared and, when differences between transcriptions reflected on-going changes, the most traditional pronunciations were retained. With assistance from Pr. Feng Wang at Peking University, the entries of the corpus with no corresponding transcription in the dictionaries were discarded, which reduced the size of the wordlist from 8 531 to 5 713. The corpus of spontaneous Japanese (NINJAL, 2011) provided transcriptions in katakana, which were then converted into phonological transcriptions by using a list of phonemic entities corresponding with morae in katakana. The initial corpus for Korean was retrieved from the Leipzig Corpus Collection and was converted into IPA by using a Korean pronunciation dictionary (Kim et al., 1993).

[3] There is also a third category which encompasses languages that express in one word what the other languages would distribute over several lexemes. These languages, such as Algonquian languages in North America, are called polysynthetic.

The WebCelex corpora in English and German (Max Planck Institute for Psycholinguistics, 2013, 2014) were retrieved online. They included an automatic transcription derived from grapheme-to-phoneme conversion as well as corresponding lemma and grammatical category for each entry of the corpus. For French, Lexique 3.80 (New et al., 2001) was used, which is very similar to WebCelex with transcription, lemma and grammatical category for each word-form of the data. In some French variants, the opposition between /e/ and /ɛ/ tends to be neutralized (Gess, Lyche, & Meisenburg, 2012) but we decided to keep those phonemes apart in the data transcription.

For Italian, the corpus PAISÀ (Lyding et al., 2014) was retrieved online and was transcribed into IPA by using the dictionary of Italian pronunciation (Canepari, 2009). When there were missing entries in the dictionary, an automatic phonemic converter (Carnevali, 2009) was used and resulting transcriptions were corrected by the first author in order to follow the transcription rules of the pronunciation dictionary. The initial corpus provided corresponding lemma and grammatical information. Swahili data were collected at the Dynamique Du Langage Laboratory (Gelas, Besacier, & Pellegrino, 2012) and lemmatized with TreeTagger (Schmid, 1995).

For FL calculation, the 20 000 most frequent word-forms and lemmas were taken into account respectively from inflected and lemmatized data in each language except for Italian with 14 629 inflected word-forms (corresponding to 8 028 lemmas) and Cantonese with 5 172 entries (due to the relatively small corpus). All phonological entries in each language were syllabified and syllabic boundaries were considered for the computation of FL. In Section 3, the influence of the following parameters was assessed: TOKEN vs. TYPE and INFlected vs. LEMmatized, which resulted in 4 potential configurations - INF/TOKEN, INF/TYPE, LEM/TOKEN, and LEM/TYPE. For each version, FL_E and *#MP* were computed for vowel and consonant contrasts. Appendix 2 provides a toy example to illustrate these different configurations. In Section 4, the FL carried by each individual vowel and consonant was calculated and discussed.

Among the four potential configurations above, the three most interesting ones will be reported in the paper. LEM/TYPE is the most lexicon-oriented dataset as it is reduced to lemmas and can be considered as a kind of "core" lexicon. On the contrary, INF/TOKEN version of data was the most usage-oriented corpus. Finally, INF/TYPE data can be regarded as the extended version of the mental lexicon. These three configurations gave insights on the structure of the core lexicon (LEM/TYPE), the influence of the inflectional morphology (INF/TYPE), and finally, the impact of the actual usage (INF/TOKEN).

3. Distribution of FL for Subsystems of the Phonological Inventory

In this section, the relative FL of each phonological subsystem (vowels, consonants, stress, and tones) are first explored in nine languages (Cantonese, English, French, German, Italian, Japanese, Korean, Mandarin, and Swahili). Further investigations are then performed with five languages (English, French, German, Italian, and Swahili) for which distinctions in terms of TOKEN/TYPE and LEMmatized/INFlected forms could be made. First, the range of variation of segmental FL is explored in the various configurations. The weights assumed by vocalic and consonantal subsystems are then examined.

3.1 Contributions of Phonological Subsystems to FL

To compute the FL of the phonological subsystems, the INF/TOKEN configuration was considered, as it was the only one available for all languages. Table 2 represents the FL associated with each phonological subsystem – vowels (FL_V) and consonants (FL_C) – as well as tones (FL_T) in Cantonese and Mandarin and lexical stresses (FL_S) in English, German, and Italian. FL reflects the relative importance of subsystem within each language.

Although the difference between consonantal and vocalic weight may be limited (as in French), FL_C was higher than FL_V in all nine languages. This result might be expected because of a universal trend to have more consonants than vowels in most of the world's languages: In LAPSyD (Maddieson et al., 2013) 646 out of 696 languages have strictly more consonants than

vowels. However, in the case of German, there were more vowels than consonants in the phonological inventory (32 vowels vs. 25 consonants in the data description) and the gap between FL_V and FL_C did not remarkably differ from those in other languages. Furthermore, the FL_v of German was the median in the dataset while the size of its vowel inventory was the largest.

While further investigating the influence of inventory size, a positive significant correlation between the size of the consonant inventory and FL_C was revealed (Spearman's $\rho = 0.792$*; *p*-value = 0.011; N = 9). There was however no correlation between FL_V and the size of vowel inventory (Spearman's $\rho = 0.519$; *p*-value = 0.152; N = 9). For instance, the FL_V of a 5-vowel language (Swahili) and that of a 32-vowel language (German) were very similar while the FL_C of Swahili with 30 consonants differed considerably from that of Japanese with 16 consonants.

The impact of lexical tone was visible, with FL_T close to FL_V in Cantonese and superior to FL_V in Mandarin. Lexical stress had also some impact in Italian ($FL_S = 0.24\%$), but almost no impact in English and German[4].

Table 2. *Functional Loads carried by vowels, consonants, tones and stress and Infra-syllabic FL_{VCTS}.*

	Languages								
	yue	eng	fra	deu	ita	jpn	kor	cmn	swh
FL_V	4.55	6.70	14.83	4.37	7.61	3.76	3.30	3.24	4.11
FL_C	10.64	20.82	19.41	15.45	11.12	9.39	11.50	13.09	20.0
FL_S/FL_T	4.48	0.005	-	0.01	0.24	-	-	4.13	-
FL_{VCTS}	62.50	52.30	55.35	47.95	44.74	44.08	45.32	58.08	53.97

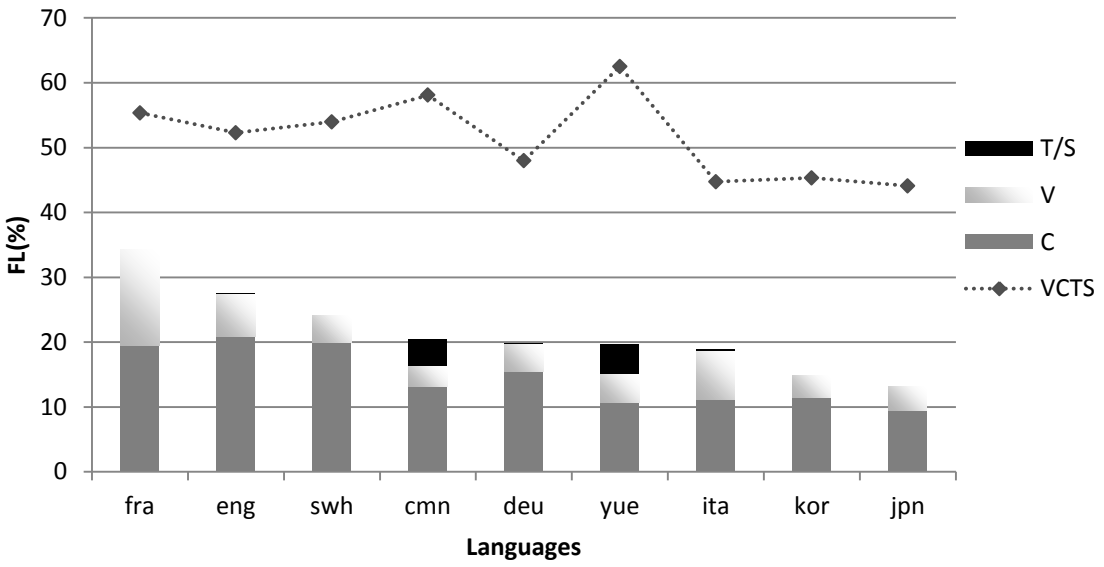

Figure 2. *Functional Loads carried by vowels (V), consonants (C), tones (T) and stress (S) and Infra-syllabic FL (FL_{VCTS}). X-axis shows languages by decreasing order of summed FL_V and FL_C.*

Information gathered in Table 2 is illustrated in Figure 2. The individual contribution of each phonological subsystem is displayed by the bars and the infra-syllabic FL_{VCTS} is represented by diamonds. Several studies have examined the relative importance of tone within a phonological

[4] In English and in German, homophony induced by stress coalescence is rare because of the high redundancy between stress and vowel quality encoding in WebCelex. Moreover when homophony arises, it impacts low frequency items.

system (Hua & Dodd, 2002; Oh et al., 2013; Surendran & Levow, 2004). Hua and Dodd (2002) highlighted that in early language acquisition, tones are acquired earlier than other elements of syllables and that their role in distinguishing lexical meaning is more crucial than phonemes. In a corpus-based study, Surendran and Levow (2004) showed that the amount of information carried by tones is as important as the amount carried by vowels in Mandarin. Oh and colleagues (2013) later confirmed this result with Cantonese data. Our results were in line with this and also suggested that there is no compensation between consonantal and tonal subsystems (see Maddieson, 2007, and Hombert, Ohala, & Ewan, 1979, for a diachronic perspective). We indeed found that both Cantonese and Mandarin relied on higher infra-syllabic FL_{VCTS} values than the other seven languages. However, the fact that the two tonal languages considered here are also isolating prevented us from concluding on the origin of the heavy weight of the infra-syllabic information. More languages, with various tone systems, would be necessary to further assess this pattern.

3.2 Frequency, Morphology, and FL

For English, French, German, Italian, and Swahili, lemmas corresponding to inflected forms were available, and INF/TOKEN, INF/TYPE and LEM/TYPE corpora could be extracted and investigated. None of these languages had tones, and lexical stress in English, German, and Italian was ignored given its very low FL with respect to consonants and vowels.

The importance of the whole phonological inventory was assessed by examining FL_{VC} (Figure 3). Cross-language variations were visible, with a similar magnitude in the three corpus configurations. For LEM/TYPE corpora, the segmental FL varied from 37.9% in German to 57.6% in Swahili. In English, German, and Italian, segmental FL was lower than 50%, which implies that distinctions between lemmas mostly relied on the structural information in these three languages. Considering inflected forms rather than lemmas (LEM/TYPE vs. INF/TYPE comparison) had a limited impact on the load carried by segments, except in Italian. However, interpretations may differ across languages. In English, the identical FL_{VC} values reflected the limited productivity of the inflectional morphology. In German (and to a lesser extent in French and Swahili), the relative steadiness observed meant that the inflectional system is relatively neutral vis-à-vis the proportion of information based upon segments. In Italian, by contrast, word-forms were more distinguished via segmental differences in the INF/TYPE configuration than in the LEM/TYPE configuration (46.0% vs. 39.7% for FL_{VC}). This result is compatible with the regular inflectional system that produces a lot of (vowel) alternations in suffixes, both in verbal and nominal morphology.

FL consequently revealed that about one half of the words' "identity" was carried by other means than segmental distinctions in these five languages. This result may reflect a balance between time-localized (i.e. segmental) information and information spread along the whole word in speech communication. Such a syntagmatic organization may be more robust to noise and local degradation than a system where most of the information on word identity depends on a short-time window. Speakers tend to modulate their utterances during speech communication in order to optimize their transmission capacity. They are also likely to reduce words with less information (i.e. words with higher predictability) by employing both surface and structural information for estimating the predictability of words (see Levy and Jaeger, 2007, among many others).

The importance of token frequency is abundantly described in psycholinguistics, where frequency effects are well documented, and it is also a corner stone in exemplar models in phonology (Johnson, 1996; Pierrehumbert, 2001, 2003; Walsh et al., 2010). Here, we looked at the global changes induced in FL patterns when comparing type and token frequencies. Although the range of the cross-language variations was almost unchanged (FL_{VC} ranged from 40.3% in Italian to 55.4% in French), language-specific effects were visible.

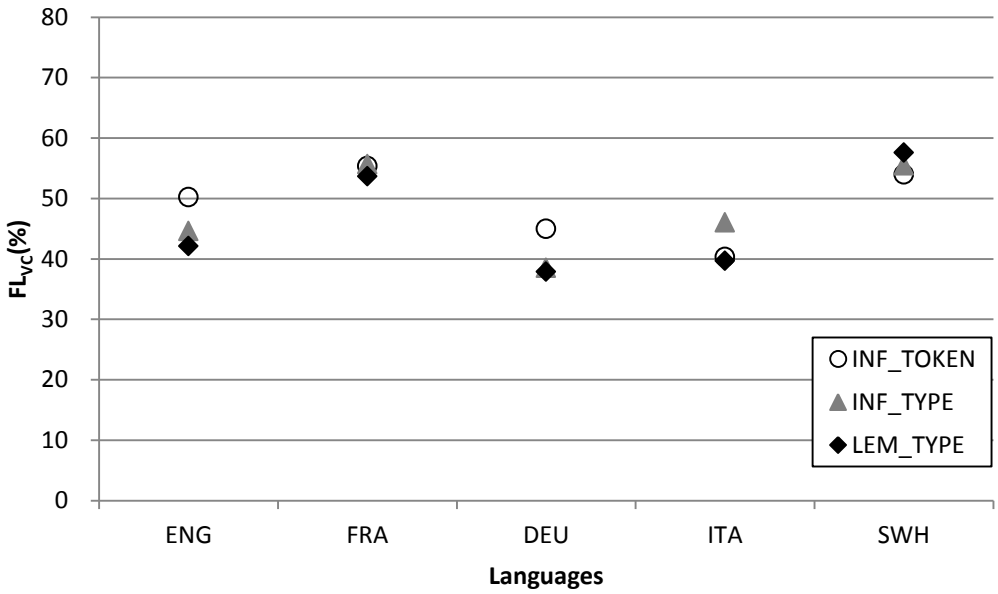

Figure 3. *Segmental Functional Load (FL_{VC}) in five languages according to corpus configuration.*

In English and German, shifting from type to token increased the weight of segments in distinguishing among inflected forms (+5.6 points and +6.4 points respectively). This effect was probably a consequence of the predominance of shorter words, including many monosyllabic words[5] in the most frequent words (Zipf, 1949; Bell et al., 2009). These words have more phonological neighbors with high frequency and they more heavily rely on segmental contrasts than longer low-frequency words since they incorporate much less structural information. An opposite trend was visible in Italian, since the segmental FL diminished from 46.0% to 40.3% from the type to the token-based corpus.

Compared to English and German, Italian has a lower syllabic complexity which clearly limits the number of monosyllabic word-forms (less than 500 are present in the corpus) and may explain this different behavior. In French and Swahili, changes induced by taking inflections and token frequencies into account were limited compared to other languages. Moreover, in the three corpus configurations, segmental loads were higher than in the other languages (values between 53.9% and 57.6% in Swahili, and between 53.7% and 55.7% in French). In Swahili, this preponderance shall be put in perspective with both the vastly predominant CV syllable structure (except in loanwords) and the strict morphological structure induced by Bantu case marking and verbal morphology. As a consequence, structural information is more limited in Swahili than in fusional languages which allow more variations, in frequent as well as infrequent word-forms. In French, the interpretation is different. On the one hand, a large variety of syllabic structures are present, allowing a large number of monosyllabic word-forms for instance (more than 3 600 are present in the corpus), in contrast to Italian and Swahili. On the other hand, the role of segments in lexical distinctions (as illustrated through the LEM/TYPE configuration) is much larger than in English and German.

An interim conclusion is that variations were visible in i) the relative weight of segmental vs. structural information in lexical distinctions and ii) the impact of token frequencies on this balance. The small language sample prevented from drawing any typological conclusions, but it suggested that the relative weight of segmental vs. structural information results from an interaction of factors that cannot be reduced to the basic size of the phonological system.

[5] The English corpus includes more than 5 700 different monosyllabic word-forms, and the German corpus more than 1 600 ones.

FL_V and FL_C values for each corpus configuration are presented in Table 3. *#MP* are not reported because of their similarity with FL_E estimated from types. FL_V ranged from 1.4% to 14.8%, whether accounting for frequency or morphology. FL_C ranged accordingly from 9.5% to 24.4%. FL_E values for INF/TYPE and INF/TOKEN configurations were highly correlated (Spearman's ρ = 0.952**; *p*-value < 0.001; V and C series pooled together; N = 10).

Table 3. *Functional Loads (in %) associated with vowel and consonant inventories, as a function of the corpus configuration in five languages (see text for details).*

			Languages				
			eng	fra	deu	ita	swh
TYPE	INF	FL_V	3.5	7.6	2.0	6.1	3.6
		FL_C	18.0	15.7	11.8	11.2	16.8
	LEM	FL_V	3.0	5.2	1.4	1.8	5.6
		FL_C	14.8	15.2	9.8	9.5	24.4
TOKEN	INF	FL_V	6.7	14.8	4.4	7.6	4.1
		FL_C	20.8	19.4	15.4	11.1	20.0

Reinforcing observations made in Section 3.1, FL_C was higher than FL_V in the five languages, for each corpus configuration. While there was a positive significant correlation between the size of the consonant inventory and FL_C for nine languages, there was none between the size of a phonological system (i.e. vowel or consonant subsystem) and its global FL neither in INF/TYPE (Spearman's ρ = 0.215; *p*-value = 0.551; V and C series pooled together; N = 10) nor in INF/TOKEN (Spearman's ρ = 0.325; *p*-value = 0.359; N = 10). These results indicated that the size of a phonological system was not a good predictor of the amount of lexical information its segmental contrasts accounted for.

3.3 Consonantal Bias

In order to investigate more specifically the potential bias towards consonants vs. vowels, we defined the difference-over-sum of FL_C and FL_V, expressed as a percentage:

$$CBias = 100 * \frac{FL_C - FL_V}{FL_C + FL_V} \qquad (8)$$

If the vocalic and consonantal subsystems have equal FL (unbiased system), *CBias* is equal to zero. The more a system is biased towards consonants, the higher *CBias* is, up to a theoretical limit of 100%. On the contrary, a system biased towards vowels would yield negative values, with a theoretical limit of –100%. The difference-over-sum provides a normalized criterion to contrast languages with each other and it is more appropriate than the difference FL_c-FL_v since a significant range of variation exists for both FL_c and FL_v.

CBias indices are given in Figure 4. Three series, corresponding to each corpus configuration, are displayed. In LEM/TYPE configuration, a strong positive *CBias* was visible for each language. It ranged from 49.1% in French to 75.2% in German.

We then explored the influence of corpus configuration (in terms of TOKEN vs. TYPE, and LEMmatized vs. INFlected data) on *CBias*. Regarding the influence of inflectional morphology, several patterns were visible on the INF/TYPE series (Figure 4). Though German and English are quite distinct from each other in terms of richness of inflectional morphology (both verbal and nominal), they exhibited almost similar patterns, with a limited impact with regard to the lemmatized configuration. In French and Italian, on the contrary, changes were notable, with *CBias* dropping from 68.2% (LEM/TYPE) to 30.0% (INF/TYPE) in Italian. In Swahili, changes between LEM and INF corpora were limited. These results suggested that this bias is not only a matter of morphological productivity.

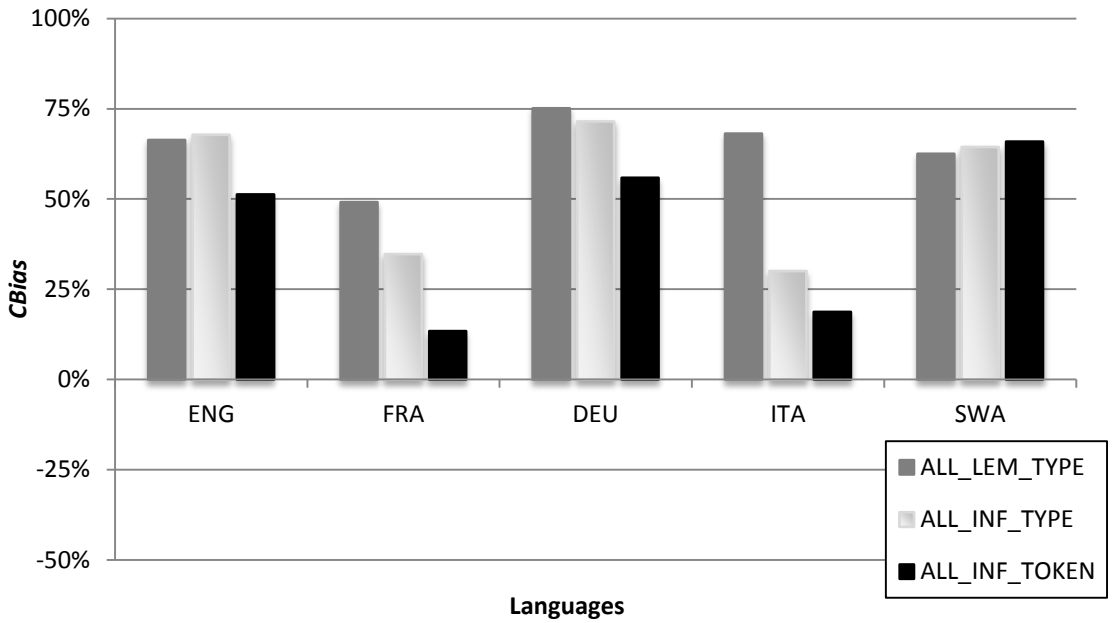

Figure 4. *CBias according to corpus configuration.*

Taking token frequency into account (INF/TOKEN series) led to decreasing *CBias*, except in Swahili. Even if it resulted in a low consonantal bias in French (13.4%), no language reached a situation biased toward vowels or even balanced. Cross-language differences were nevertheless much more visible in this configuration than in the LEM/TYPE configuration previously discussed, with *CBias* ranging from 13.4% in French to 65.9% in Swahili.

This approach revealed the existence of a large *CBias* in the core lexicon (LEM/TYPE configuration) in the five languages. The magnitude of this effect was not directly linked either to the absolute size of the vowel system (Swahili exhibited a large value with a 5-vowel system) or to its relative size compared to the number of consonants (German showed the highest *CBias* though it has more vowels than consonants). Moreover, *Cbias* seemed to be insensitive to syllabic complexity and syllable inventory size (English and Swahili reached similar magnitudes with very different syllabic complexities). The comparison of LEM/TYPE and INF/TYPE configurations provided a way to evaluate the impact of the inflectional morphology. Two profiles were shown. On the one hand, morphology had a limited impact on *CBias* in English, German, and Swahili, though these languages drastically differ in their morphological productivity. On the other hand, inflectional morphemes had a tendency to counter-balance the bias towards consonants in French and especially in Italian. Finally, when token frequency is considered, i.e. when we switched from a "flat" lexical representation of word-forms to a usage-based representation, the *CBias* range of variation became larger, even if this pattern was still present in the five languages.

Computing the *CBias* for Cantonese, Japanese, Korean, and Mandarin in INF/TOKEN configuration led to 40.1%, 42.8%, 55.4% and 60.3% respectively. These values were all positive, and fell within the range of previous values.

These results suggested that the consonantal bias may be a robust trend at the lexical level, beyond large typological differences among languages in terms of size of phonological system, syllabic complexity, and morphology. This *CBias* was nevertheless modulated by usage, with possible consequences on the cognitive representations of the speakers.

4. Distribution of FL within Phonological Subsystems

In this section, all nine languages are considered in INF/TOKEN configuration. The distributions of FL_E and *#MP* are investigated in the vowel and consonant subsystems, as well as their consequences in terms of system economy. The individual phonemes with the highest

FL_E and *#MP* in each language are then discussed from a typological perspective. Like in Section 3, the 20 000 most frequent word-forms were employed, except in Cantonese and Italian where only 5 172 and 14 629 entries were present respectively, due to limitations in corpus size. Language data and preprocessing were previously described in detail in subsection 2.3.

4.1 Patterns in FL Distributions

Up to this point, we presented cumulative results, at the scale of each phonological subsystem or at the more global scale of infra-syllabic information as a whole. FL is also useful to rank contrasts within a language subsystem and to cross-linguistically compare their distributions. In Figures 5 and 6, such distributions are displayed for vowels and consonants respectively. Pairs are ranked by decreasing order of FL on the x-axis with FL_E on the left y-axis (grey triangles) and *#MP* on the right y-axis (black circles). Since the number of contrasts lawfully followed the number of vowels and consonants in each language according to a $n(n\text{-}1)/2$ relationship, x-axis ranges differ between languages. Accordingly, the y-axes depend on FL_E and *#MP* values but scales have been matched in order to ease comparison of the distribution shapes. The first striking observation is that none of the nine languages evenly relied on its vowel or consonantal system to carry its FL. For both vocalic and consonantal contrasts (for *#MP* and FL_E), the general shape consisted of two sections: high-ranked contrasts, characterized by a rather abrupt decline, and low-ranked contrasts, with a slow decrease. The relative size of each section might be variable, but most of the time, it consisted of five pairs or less, which is a very small number of contrasts to rely on. Despite this common trend towards uneven distributions, language-specific differences were also visible. In some cases, the decline was regular, without any clear inflection point (e.g. distribution of vowel contrasts in German or Cantonese, or distribution of consonant contrasts in English). On the contrary, Italian for vowels and Japanese for consonants exhibited "S-shape" distributions. In Italian, the first two vocalic contrasts were involved in almost the same number of minimal pairs, and the same pattern held for consonants. In other cases, the decrease in FL between the first and the second contrast was large (e.g. in Japanese, Korean, Swahili for vowels and in German and Korean for consonants). Cross-linguistically, phonological contrasts didn't follow a regular distribution, such as Zipf's law (observed for word-form frequencies) or another heavy-tailed distribution (such as Yule distribution, see Martindale et al., 1996).

Comparison between *#MP* and FL_E distributions may also be insightful since they point towards potentially different cognitive processes. *#MP* distribution is related to the whole set of word-forms in the language, and it thus corresponds to the organization of the mental lexicon. In contrast, by including token frequency, FL_E is more related to frequent words and to online processing in situations of communication. In several cases, the two distributions were analogous (e.g. Korean for consonants, or Mandarin for both vowels and consonants). In other cases, the different distributions observed meant that the structure of the basic lexicon (consisting of the frequent word-forms) differed from the structure of the extended lexicon. More precisely, two patterns were present. When FL_E distribution was partly above the *#MP* distribution, as for vowels in Korean or consonants in German or Swahili, a few contrasts were promoted by usage. On the contrary, having the FL_E distribution below the *#MP* distribution signified that for frequent word-forms, less information was conveyed by the infra-syllabic level. This pattern is common in our sample (in German, English, Italian, and Cantonese for vowels and in French, Japanese, and Cantonese for consonants). It may be related to the amount of other linguistic information available, which helps to understand words and consequently limits the burden carried by each word itself.

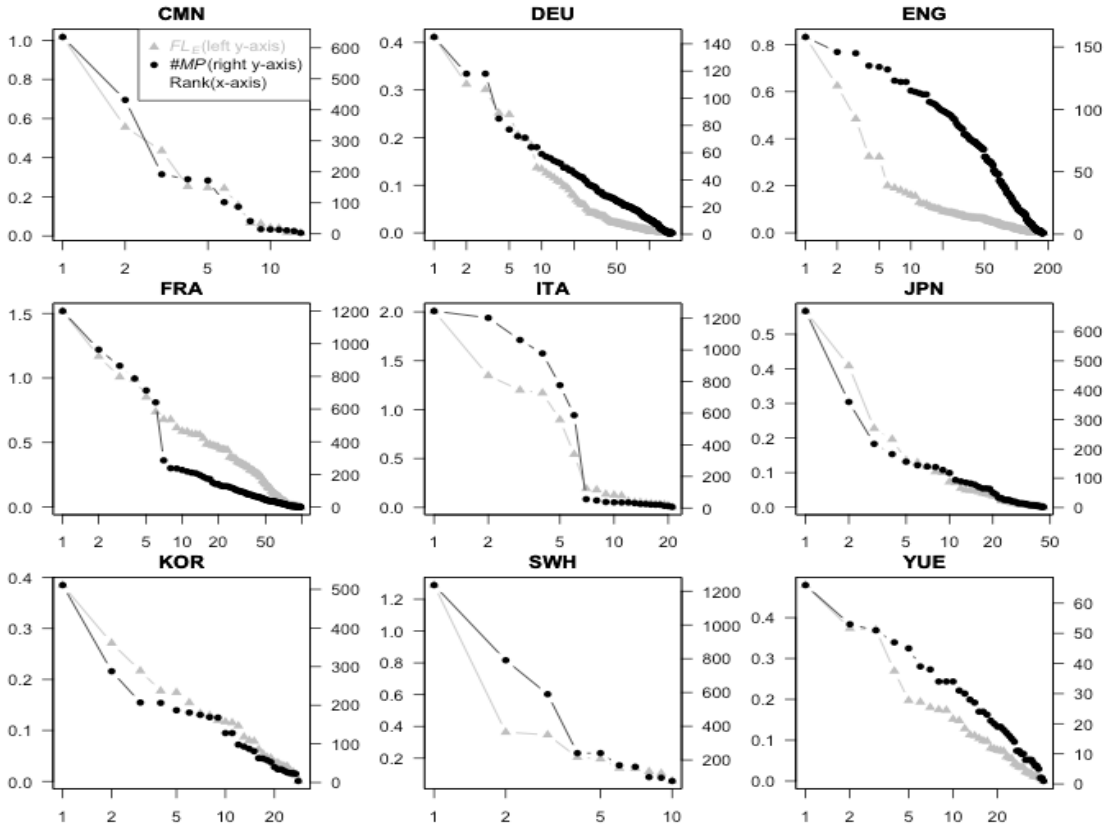

Figure 5. *Distribution of Vowel Pairs: FL_E on the left y-axis (in gray) and #MP on the right y-axis (in black). Pairs are listed by their decreasing order of FL values using a logarithmic scale*.

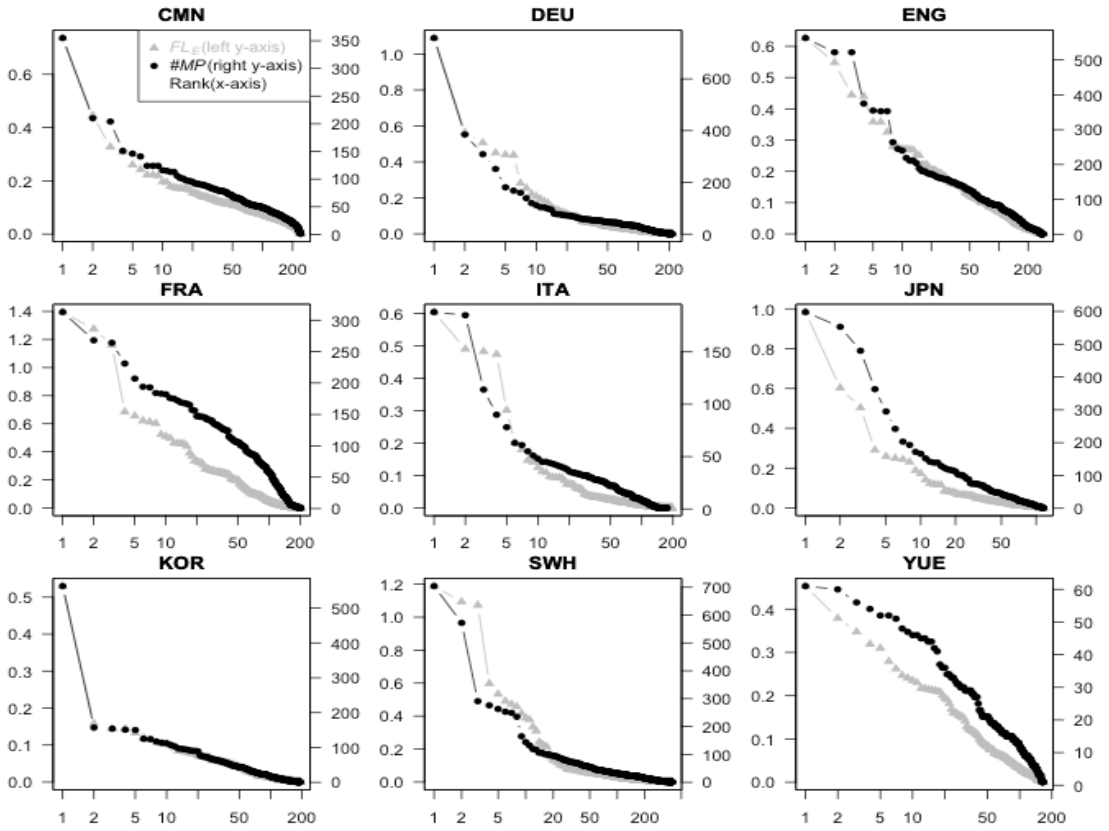

Figure 6. *Distribution of Consonant Pairs: FL_E on the left y-axis (in gray) and #MP on the right y-axis (in black). Pairs are listed by their decreasing order of FL values using a logarithmic scale.*

We showed in Figures 5 and 6 that a lot of contrasts were characterized by a very low FL and that they marginally contributed to the segmental FL. They conveyed consequently a very low amount of information and we performed a simulation in order to evaluate how the nine languages behave at the systemic level in this respect. The algorithmic principle was to reduce the phonological set, by iteratively eliminating the segment with the smallest FL until only one segment remained. For instance, in Swahili, we observed for the vowels: FL(/e/) < FL(/o/) < FL(/u/) < FL(/i/) < FL(/a/). In the first iteration, /e/ was eliminated from the system, and coalesced with the vowel /a/ with which it was involved in the maximum number of minimal pairs. We computed the relative loss of entropy corresponding to the lexicon described by this new 4-vowel system. In the second iteration, /o/ underwent the coalescence process, resulting in a lexicon described by a 3-vowel system. The process was next applied to /u/, then to /i/, and resulted in a 1-vowel system (with entropy consequently equal to FL_V). The results of this simulation are displayed in Figures 7 and 8. For legibility, the y-axis represents the proportion of initial entropy preserved in the altered system. It is thus the complement of FL on 100%. The iteration step in the simulation is indicated on the x-axis (zero being the original system, with a FL of 100%).

Two major patterns are visible in the graphs. The first configuration illustrated that some systems were more sensitive to changes induced by the reduction process. This pattern was present for instance in Korean and Swahili for vowels, and in Mandarin, Japanese, and Cantonese for consonants. In most cases, however, systems were very resilient to reducing the size of the phonological systems, and the loss in FL induced was barely noticeable at least at the beginning of the process. It was especially salient in German and English for vowels and for German, English, French, Italian, Korean, and Swahili for consonants. In German, for instance, the majority of the vowel system could be coalesced with an information loss of less than 1%.

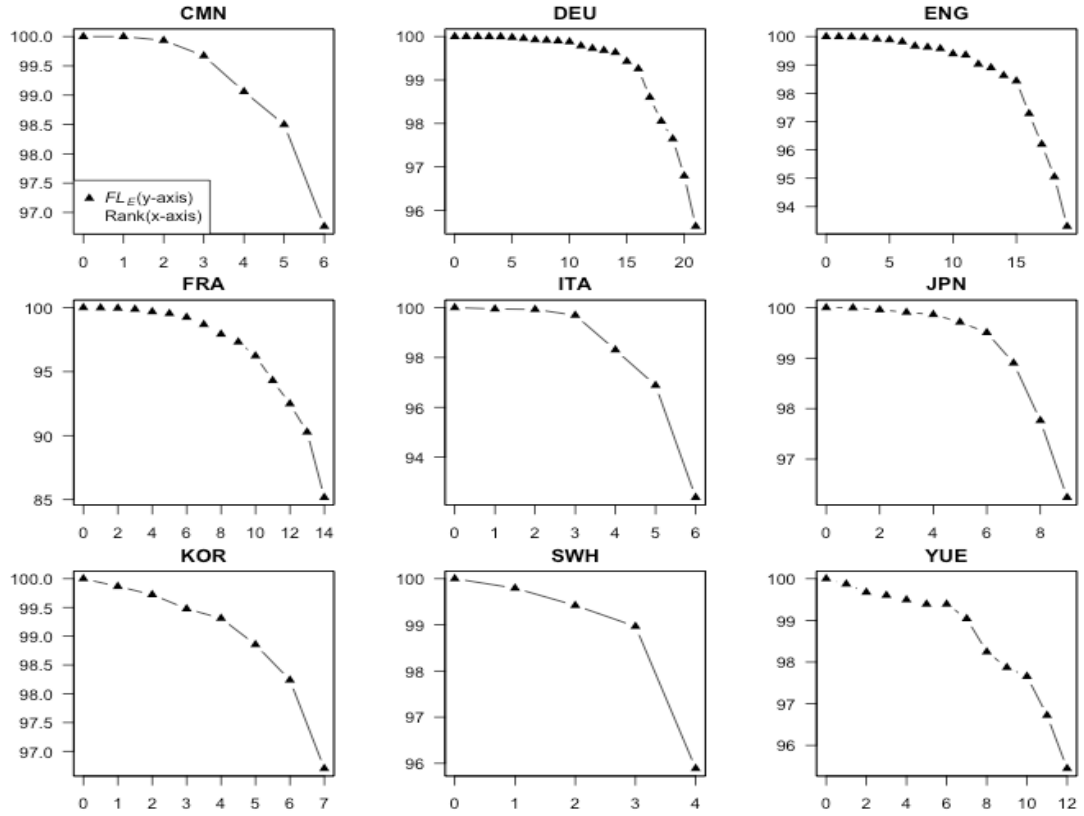

Figure 7. *Simulation of the Relative Loss of Entropy Induced by Reducing Vowel System, % of FL_E on the y-axis (in black), phonemes listed by their increasing order of FL (x-axis).*

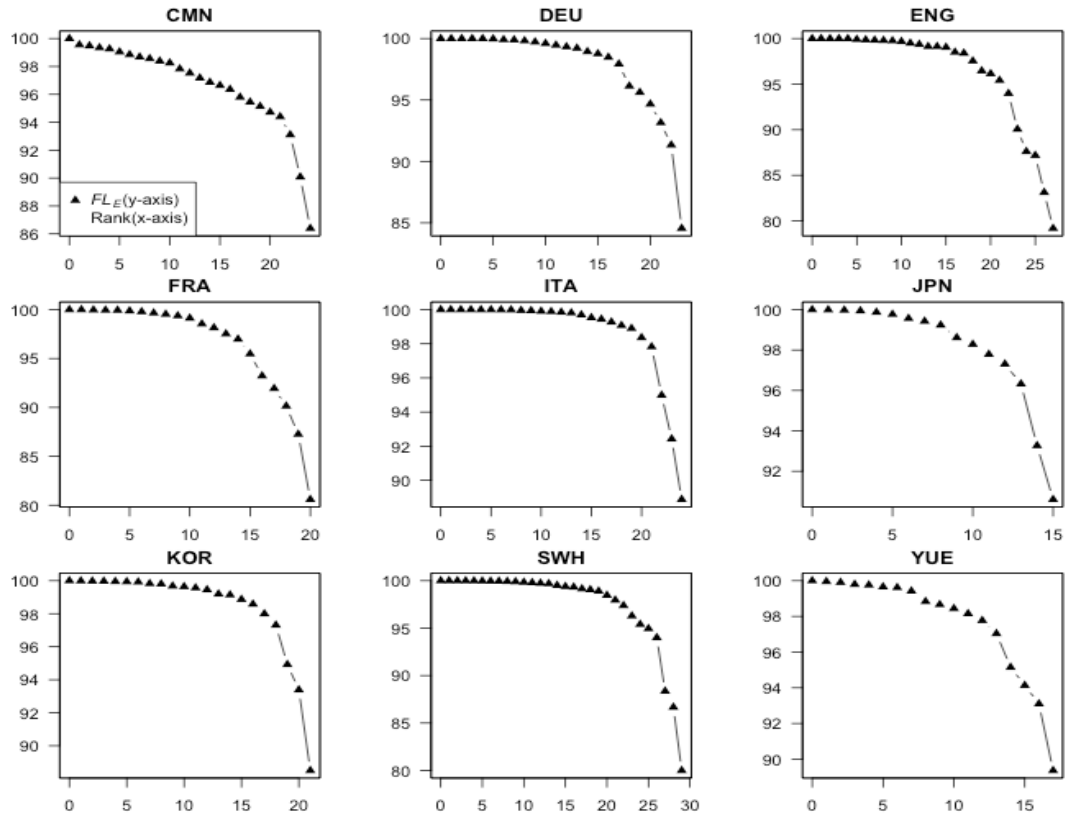

Figure 8. *Simulation of the Relative Loss of Entropy Induced by Reducing Consonant System, % of FL_E on the y-axis (in black), phonemes listed by their increasing order of FL (x-axis).*

In general, the amount of information loss induced by a merger varied more widely in consonants than in vowels among the languages, which is consistent with the larger FL associated to the consonantal component. One major exception is Italian, for which a drastic reduction of the number of consonants would have minor consequences in terms of information loss. This result is coherent with Section 3 on *CBias* where the importance of consonants in structuring the lexicon was highlighted, but limited in Italian.

One could consider that keeping such contrasts distinct in a language is costly and provides no real advantage, especially if these contrasts rely on segments that don't participate in any high FL pair. However, Vitevitch (2008) described the self-organization of phonological word-forms in the mental lexicon by employing the concepts of small-world topology and scale-free network. These networks are characterized by a small average path length, a high clustering coefficient (for both network patterns), a power-law degree-distribution and a preferential attachment (for the latter) in the growth theory of Barabási-Albert (Barabási & Albert, 1999). In this approach, the structures of phonological system and mental lexicon can both be described as scale-free networks due to their preferential attachment - a small number of giant components with many other smaller components. Such properties facilitate language acquisition, production and perception with its robustness and resilience to errors and damages of components. From this perspective, the observed distribution of vowel and consonant systems shown in the figures above can be regarded as the consequences of cognitive efficiency and optimization for language acquisition and information retrieval, which is a robust property of natural languages. For instance, Morales and Taylor (2007) have shown that variable frequencies of language elements improve language acquisition compared to the elements with equal frequencies. Such characteristics of a natural language which self-organizes the structure of its systems result from the cognitive efficiency and optimization during language learning and speech communication.

4.2 Cross-language Trends in Preferred Phonological Features

Figures 5 and 6 pointed towards the high proportion of information coded by the five highest-ranked contrasts in the nine languages. Consequently, we further examined these specific contrasts in this subsection, as well as the highest-ranked segments themselves. Tables 4 and 5 display the five vowel pairs with the highest FL_E computed with the INF/TOKEN configuration of corpus and the five individual vowels with the highest FL_E respectively.

Among the five vowel pairs with the highest FL_E listed by their decreasing order of FL_E in Table 4, there was no pair which was present in all the nine language studied. In fact, we observed 28 different contrasts (the maximum possible being 45) composed of 18 different vowels. However, four contrasts appeared in four different languages: /i-a/, /i-u/, /e-a/ and /o-a/. Interestingly, they rely on /i, e, a, o, u/, the five most frequent vowels in the world's languages. Among those four contrasts, three involved the low vowel /a/, this vowel being implicated as well in eight of the nine most important contrasts found in our sample. This points towards a particular role of the maximally open vowel. The only language without the vowel /a/ in its most salient contrast is Korean, with the pair /i-e/. This time it's the maximally closed vowel that is found. Again vowel height seems to be an important dimension for vowel oppositions as it operates in 16 out of the 28 different most salient contrasts, either maximally /i-a/ or minimally /i-e/ for example.

Although Swahili obeyed a kind of maximum contrast selection (with respectively /i-a/, /u-i/ and /u-a/ on the podium), the general trend was to prefer moderate to low acoustical distances in these contrast sets, as illustrated by /ɔː-aː/ in Cantonese or /a-ɛ/ in German. Redundant contrasts, defined as contrasts where more than one feature (frontness, aperture, and rounding) is involved, were also very common but they were rarely based on a secondary feature, with the exceptions of /aː-ɐ/ in Cantonese and /ɑ̃-e/ in French. In Italian, three of the five pairs with the highest FL_E, (/e-a/, /i-e/, and /i-a/) seem to reflect the inflectional morphology as they contain the thematic vowels /a/, /e/, and /i/, which is the marker of inflection class in verbal morphology (Da Tos, 2013).

	Languages																	
	yue		eng		fra		deu		ita		jpn		kor		cmn		Swh	
1	ɔː-aː	0.48	aɪ-eɪ	0.83	e-a	1.52	a-ɛ	0.41	e-a	2.01	e-a	0.57	i-e	0.39	ə-a	1.02	i-a	1.29
2	ɛː-ɔː	0.37	ɪ-æ	0.62	ø-e	1.17	ɪ-aɪ	0.31	i-e	1.35	o-a	0.41	o-i	0.27	u-i	0.56	u-i	0.36
3	o-ɐ	0.37	eɪ-iː	0.48	ø-a	1.01	a-ɪ	0.30	i-a	1.20	i-a	0.23	i-a	0.22	u-ə	0.44	u-a	0.35
4	aː-ɐ	0.27	aɪ-iː	0.32	ɑ̃-e	0.99	aː-iː	0.25	o-a	1.17	o-e	0.20	o-e	0.18	u-a	0.25	e-a	0.21
5	u-i	0.20	ɪ-ɒ	0.32	ɛ-e	0.85	a-aɪ	0.25	o-i	0.90	u-i	0.14	o-a	0.17	y-i	0.25	o-a	0.20

Table 4. *5 Vowel Pairs with the Highest FL_E*

Several remarks can be made at the level of the vowels themselves (Table 5). First, the differences among the five vowels with the highest FL_E were less important than the ones between the five most salient contrasts, this means that the load is more evenly divided at the level of the segments than what appears to be when looking directly at contrasts. Second, for almost all languages, the vowels with the highest FL_E were the ones implicated in the five most salient contrasts. When looking at the vowel qualities present in this set, we observed 24 different vowels (again maximum is 45). The low vowel (/a/-like) was not always the preferred attractor or hub, (four languages out of nine) but it was present in the table for each language, either as a monophthong or as the beginning of a diphthong. It is followed by /i/ or /iː/, present in eight out of nine languages. /e/ and /o/ or /oː/ were found in five languages. Surprisingly, the back high vowel /u/ is only present in two languages (Mandarin and Swahili), yet the five most frequent vowels are the most contrast-bearing ones. In terms of features, among the 45 vowels and diphthongs of the table, 23 vowels are front, 10 are central (incl. /a/-beginning diphthongs)

and 12 are back. Finally, we noticed that the larger the vowel inventories, the more likely the set of "preferred" vowels will be to include vowels other than /i, e, a, o, u/.

	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

Table 5. *5 Individual Vowels with the Highest FL_E*

The first remark that can be made for consonants (Tables 6 and 7) is that they show more variability than vowels. We observed 37 different contrasts out of the 45 possible relying on 22 different consonants. Only six contrasts were present in more than one language: three in three different languages and three in two different languages. All six contain coronal consonants and four include a nasal. These trends can in fact be generalized across the entire set of preferred contrasts. The first ranked contrast involved at least one coronal consonant in all 9 languages. More generally, coronal consonants are present in 43 of the 45 contrasts listed in Table 6, with a prominence of the voiced nasal /n/ (in 18 contrasts), followed by the voiced stop /d/ and the lateral approximant /l/ (both in 9 contrasts). In terms of manner of articulation, oral and nasal stops, fricatives, affricates, and approximants are present, with a preference for nasals and stops, followed by fricatives and approximants.

	Languages																	
R	yue		eng		fra		deu		ita		jpn		kor		cmn		swh	
1	n-m	0.45	n-t	0.63	l-d	1.40	R,r-n	1.09	l-n	0.60	s-k	0.98	l-n	0.53	t-l	0.74	j-n	1.19
2	ts-t	0.38	z-t	0.55	l-s	1.28	R,r-m	0.57	s-d	0.49	w-g	0.60	g-t	0.16	ŋ-n	0.45	j-w	1.09
3	ts-k	0.35	h-ð	0.44	s-d	1.16	z-d	0.51	l-d	0.48	n-t	0.50	n-g	0.14	t-ʂ	0.33	w-n	1.07
4	ts-j	0.32	n-z	0.44	n-d	0.69	s-n	0.45	n-d	0.47	m-n	0.29	n-d	0.14	tʂ-k	0.31	z-j	0.60
5	ts-s	0.31	ð-b	0.36	l-n	0.66	v-d	0.44	k-l	0.30	m-k	0.26	n-m	0.13	tɕ-ɕ	0.26	j-l	0.53

Table 6. *5 Consonant Pairs with the Highest FL_E*

Table 7 shows the five consonants with the highest FL_E. We found 19 different consonants out of the 45 possible, 13 of which were coronal. 8 out of 19 different consonants were found in more than one language. Only two of them were not coronals (/m/ and /k/). Coronal consonants appeared with various manners in the first row in all languages except Japanese (/k/). Another general trend was a preference for voiced consonants, which accounted for 27 consonants out of all 45. This preference was nevertheless relative, since five out of nine first-ranked consonants were voiceless, and this reversed tendency even pervaded almost the entire table for Cantonese and Mandarin. In this regard, we can note that when the consonant inventory of the language includes voiced stops, the most frequent contrasts rely on sonorants, whereas if the inventory lacks voiced stops the most frequent contrasts involve obstruents.

	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	sʰ	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

Table 7. *5 Individual Consonants with the Highest FL_E*

Finally, we adopted a different perspective by investigating the FL distribution in terms of distance between the members of contrastive pairs. Figure 9 shows FL_E distributed according to a feature distance, for vowels (Left panel) and consonants (Right panel). The feature distance between two segments was computed on the basis of their segmental definitions in the UPSID database (Maddieson, 1984; Maddieson & Precoda, 1990). Specifically, features are compared within the natural classes they belong to (frontness, roundedness, manner, place, etc.), and the distance is equal to the number of classes in which segments differ. Secondary contrasts such as *nasalized* or *long* define distinct additional classes. For example, the distance between /i/ {*high*; *front*; *unrounded*} and /u/ {*high*; *back; rounded*} is 2. The distance between /oː/ {*long*; *lower-mid*; *back*; *rounded*} and /õ/ {*nasalized*; *lower-mid*; *back*; *rounded*} is 2 also since the *nasalized* and *long* features belong to two distinct classes. The distance between /p/ {*unvoiced*; *labial*; *occlusive*} and /v/ {*voiced*; *labio-dental*; *fricative*} is 3. In the data set, most distances ranged from 1 (e.g. /n-m/) to 4, with six contrasts yielding a distance of 5 ((/ãː-ʏ/, /ãː-ɔ/, /ɐ̃ː-au/ in German and /kʰʷ-j/, /w-kʰʷ/, /m-kʰʷ/ in Cantonese). For vowels in the nine languages, more than 50% of the FL was carried by distinctions of one or two features except in Cantonese and English. However 2-feature contrasts were favored over 1-feature or 3-feature contrasts, except in Cantonese, English, French, and Mandarin. It highlighted a trend to prefer redundant vocalic contrasts over the most economical ones (1-feature contrasts). In Mandarin, 1-feature contrasts almost accounted for one half of the total FL_V by themselves. On the contrary, in Cantonese, English and French, 3-feature contrasts were the most important. One can also mention that in French, 4-feature contrasts were the favored ones. They involved one nasal vowel and one oral vowel with qualities differing in their 3 dimensions and their importance may be related to the frequent use of grammatical words consisting only of one nasal vowel (such as on [ɔ̃], un [œ̃], en [ɑ̃]).

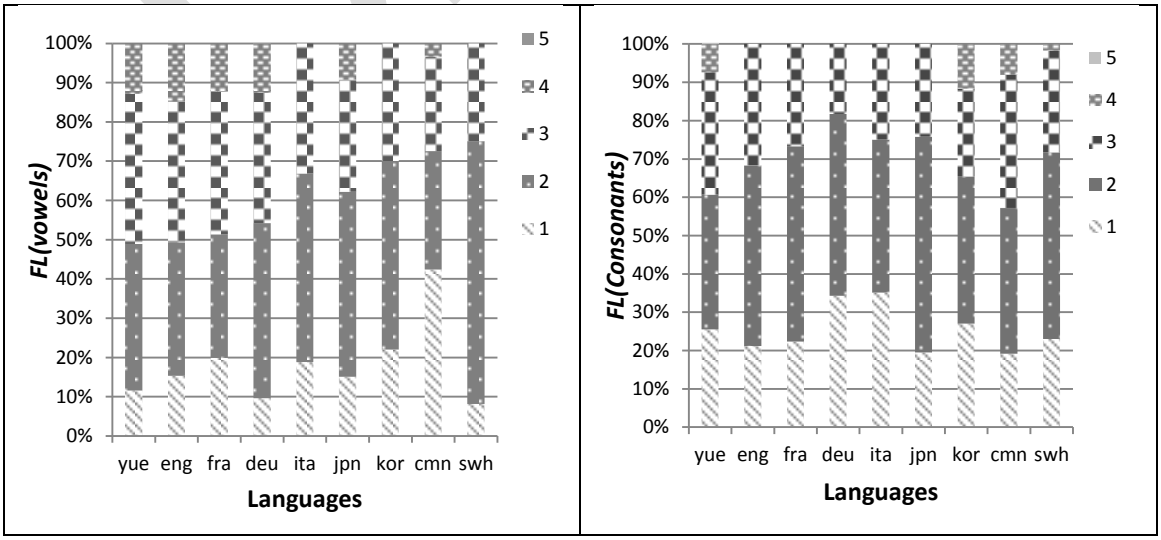

Figure 9. *Distribution of FL_E as a function of feature distances of the contrasts. Left: vowels. Right: consonants.*

For consonants, "2-feature-or-more" contrasts were in majority, and similarly to vowels, a cross-linguistic tendency towards an economical system was illustrated by the predominance of 2-feature contrasts.

Comparing the results of vowel systems and consonant systems in the 9 languages leads us to assume that cognitive principles for organizing vowel and consonant systems are different in nature. In the case of vowel systems, languages employ the principle of maximal perceptual contrasts (Jakobson, 1941) for organizing phonological structures and lexicon. In the context of language acquisition, Rose (2009), as cited in (Van Severen et al., 2012), mentioned that consonants with high FL tend to have the least articulatory complexity and the highest perceptual salience, which corresponds to the characteristics shared by the coronal consonants. Presumably, different acoustic characteristics of vowels and consonants may also play an important role regarding the different organizations of both vowel and consonant minimal pairs - the perception of consonants is more categorical and the perception of vowels is more continuous (Liberman et al., 1957, Fry et al., 1962).

5. General Discussion

The information-theoretic approach implemented in this paper directly bridged the level of the phonological components and the level of the lexicon. We thus proposed a shift from a common view of phonological systems as an inventory of components (segments, stress, tones) toward a functional perspective encompassing lexical relationship between these components. This approach relies on large corpora and facilitates cross-language comparison since the same methodology was applied to each language[6].

5.1 FL at the level of phonological subsystems

The study presented in Section 3 gave support to the existence of a lexical consonantal bias in five languages (two Romance languages, two Germanic languages, and one Bantu language). Japanese, Korean and two Sinitic languages were further examples where FL_C was much larger than FL_V. The index we defined, *CBias*, ranged from 49% to 75%, reflecting a preference toward consonant-based distinctions rather than vowel-based distinctions when analyzing a corpus of lemmatized forms and leaving token frequency aside. However, this trend was modulated as soon as inflected word-forms and/or token frequency were considered. In the INF/TOKEN corpus configuration, for instance, *CBias* ranged from 13% in French to 66% in Swahili, with various tendencies among the languages. Consequently, this consonantal trend should not be seen as an absolute and monolithic phenomenon since it resulted from the interaction between several linguistic dimensions (phonological inventories, but also syllabic diversity, and morphological type, as well as differences between lemmas and affix structures). For instance, Italian and Swahili had somewhat similar *CBias* at the lemmatized level (respectively 68% and 63%) but their behavior drastically differed in the INF/TOKEN configuration (resp. 19% and 66%). These observed differences between type and token FL may be related to language-specific configurations in phonological representations and mental lexicon, with consequences for online processing as well as on the dynamics of language acquisition[7].

[6] Additional studies will obviously be necessary to extend the report done in the following lines to a larger number of languages. We thus don't pretend reaching any typological conclusion given the small language set studied so far. Similarly, the robustness of our approach has to be more thoroughly investigated. Preliminary experiments showed that distributional patterns seem to be robust against the variation in the corpora size.

[7] For instance, Kissling (2012) showed that phonological differences in two languages impact short-term memory processing. More precisely, she showed that English native speakers recall vowel series better than consonant series whereas the reverse is true for Arabic native speakers. In our opinion, corpus

In their seminal paper, Nespor, Peña, and Mehler (2003) advocated for a greater relevance of consonants to build the lexicon, and a greater relevance of vowels to carry grammatical information, and they mentioned linguistic and cognitive motivations. They indicated the facts that most languages have more consonants than vowels in their inventories, that the number of consonantal "slots" is larger or equal to the number of vocalic slots in syllables (except in the basic V syllable structure) and finally that consonants have a general tendency to disharmonize within words, while vowel harmony (as well as vowel reduction) is frequent in the world's languages. According to these authors, these factors converge towards a more salient role of consonants than vowels in word distinctiveness[8]. Further evidence comes from psycholinguistic experiments on word transformations (Cutler et al., 2000) and later confirmation in language acquisition (Nazzi, 2005; Nazzi & New, 2007). Nespor, Peña and Mehler also mentioned that in the area of inflectional morphology, the "division of labor" between consonants and vowels has some "fuzzy boundaries", leaving a more thorough assessment to future investigation (2003:204). Nazzi and New shed some light on this issue by showing that in French the whole lexicon (roots and inflected forms) relies less heavily on consonantal contrasts than lexical roots only, when types are considered (Nazzi & New, 2007:277). They thus endorsed the influence of morphology on the relative role of consonants in the lexicon. This statement was supported by the present study, as the *CBias* effects for French and Italian indicated that the inflectional system moderates consonantal bias to some degree, in contrast with the effects for German. More generally, comparing *CBias* between LEM/TYPE and INF/TYPE configurations may help refining the "fuzzy boundaries" for each language considered.

Moreover, recent studies show that the role of consonants to access the lexicon might not be as monolithic as supposed, and especially that there is an interaction between the information carried by consonants or vowels and their position in words. Estimating this information through conditional entropy, Tanaka-Ishii established that in English, at the beginning and at the end of words, information carried by consonants is much larger than information carried by vowels, while within words, this difference is reduced (Takana-Ishii, 2012). Very recently, Delle Luche and colleagues also showed that consonantal bias is sensitive to the syllable and rhythm structure of words in French and English (Delle Luche et al., 2014). Finally, it is important to notice that the consonant advantage visible in the lexicon disappears in production and perception, and is even replaced by a vowel advantage, when whole sentences are considered (Fogerty & Humes, 2012; Kewey-Port, Burkle, & Lee, 2007; Owren & Cardillo, 2006). Stilp and Kluender (2010), in a radical acoustic approach that doesn't consider segments as primitives, also show a prevalence of vowels over consonants in speech intervals characterized by high values of their index of cochlea-scaled spectral entropy (and thus high information amount). The approach developed in this section didn't address the balance between consonantal and vocalic information in sentences since it was based on lexical data. However, the differences observed between processing at word and sentence levels are consistent with the importance of temporal organization of information in speech. Under this view, the differences in lexical structures revealed in this section, for instance between type frequency and token frequency, may reflect this prominence, since token frequency not only influences cognitive representations but also expectations (and thus information) in the processing of connected speech. The corpus-oriented study presented here, although limited, can complement other approaches, such as behavioral experiments in the search for explanations of the distinct role of consonants and vowels in language. Section 3 also aimed at studying the relative contribution of vowels, consonants, stress, and tones to lexical distinctions. The importance of tone system in Cantonese and Mandarin was first confirmed. Together with their

studies based on data collected during language acquisition would offer an interesting perspective to complement psycholinguistic experiments on vowel and consonant perception and representation.

[8] It has also been argued that speech consists more of consonantal than vocalic substance (in terms of duration), but Easterday, Timm, & Maddieson (2011) mitigated this assumption since in their corpus of 22 languages, the proportion of vocalic duration ranged from 43.3% to 60.1%, with an average of 53.8%.

isolating morphology which strikingly limits the structural information in the lexicon, it might explain the large infra-syllabic FL observed for these two languages (63% and 58% respectively). Among the nine languages on average, 51.7% of the lexical distinctions relied on infra-syllabic components. It pointed towards a balance between localized short-term information (measured by infra-syllabic FL) and longer term information. One has nevertheless to keep in mind that the phonemic transcriptions of word-forms only provide part of the picture. The speech phonetic substance is not in a one-to-one relationship with the phonemic "theoretical" sequence and continuous speech moreover involves predictability effects that alter the realization and perception of the words themselves (see Aylett & Turk, 2004; Levy & Jaeger, 2007; Piantadosi, Tily, & Gibson, 2009 for discussion).

5.2 FL Distribution within Phonological Subsystems

As developed in Section 4, uneven distributions of FL among the available contrasts were present in the nine languages and suggested the existence of a cross-linguistic trend. Hockett's diagnostic quoted in the introduction was thus confirmed, and our quantitative approach also shed light on the concentration of FL on very few contrasts (Figures 5 & 6). In the case of vocalic contrasts, they were moreover built upon a small set of vowels, while, for consonants, these high-FL contrasts are more disseminated over the consonant system, yet it is important to note the strong presence of coronals and nasals in the set of most salient consonants. Finally, we observed a small significant negative correlation between the FL of consonantal contrasts and the feature-distance of its constituents: the higher the FL, the closer the members of the pair (it was just a tendency for vowels).

Finally, a remarkable trend was illustrated in Figures 7 and 8 despite the differences in phonological inventories among the sample. FL concentration on a few contrasts also resulted in a kind of resilience of the lexicon vis-à-vis an alteration of its phonological inventory. For the nine languages, the simulations based on an iterative process of coalescence, yielded a two-phase pattern: removing step by step the majority of the phonemes led to a gradual and limited decrease of the FL. The second phase, characterized, on the contrary, by an abrupt slope , led to major changes in the information encoded by the phonological system. It would be interesting to reproduce the same methodology with a larger number of languages.

The existence of cross-language trends should not hide that language-specific patterns were also revealed. For instance, the differences between FL_E and *#MP* distributions (Figure 5 and 6) widely varied from one language to another, especially for vowels. In some cases, taking token frequency (as in FL_E) into account led to more continuous distributions while in other cases, considering only minimal pairs, without any usage-based count (as in *#MP*) yielded the most regular distributions. Such differences might i) mirror structural differences in the language lexicon and ii) have consequences on the cognitive processing of the speakers' mental lexicon. Further studies, including a more comprehensive examination of each language distribution, will be necessary to go beyond this simple report.

5.3 Conclusion

We would like to highlight that the distributions studied here may be put in relation with graph representations of lexicons, phonological systems, etc. The methodology presented here makes the phonological system *emerge* from interactions between word-forms in a lexicon. These interactions are often represented as graphs, and their regularities are often viewed as mirroring the phenomena from which they develop (see Arbesman, Strogatz & Vitevitch, 2010; Gerlach & Altmann, 2013; Jäger, 2012; Kello & Beltz, 2009; and Kello et al., 2010, for discussion). When it comes to language, emergence can be considered at different levels. Moulin-Frier et al. (this issue) emphasizes how phonological properties may emerge from a set of nonlinguistic (cognitive, motor, perceptual, communicative) abilities. Implementing language games additionally highlights how properties shared by a community of speakers may emerge from local interactions. These two perspectives are at work in the COSMO model. However, the linguistic structures manipulated in language games cannot yet approach the complexity of real word-forms, and FL is thus insightful for investigating how actual word-forms interact.

Avoiding homophony arising from phonetic change, for example in the case of the loss of stop codas /p,t,k/ between Late Middle Chinese and Standard Mandarin, may lead to the emergence of new phonemic contrasts. Moreover, other evolutions may take place, as it was the case in Chinese, at the morphological level with the disyllabification of words, which reduced homophony. Diachronic corpora of texts may therefore be useful to test evolutionary hypotheses, and move beyond synchronic analyses of FL as those performed in this paper.

Acknowledgements

The authors are grateful to the LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon for its financial support within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French government operated by the National Research Agency (ANR). The authors also thank Hadrien Gelas and Gérard Philippson for their help with Swahili data collection and expertise, Sébastien Flavier and Ian Maddieson for their help and guidance with LAPSyD data.

References

Aikhenvald, A. Y. (2007). Typological distinctions in word-formation. In T. Shopen (Ed.), *Language Typology and Syntactic Description, Vol. 3*. Cambridge: Cambridge University Press, 1-65.

Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, 20(03), 679-685.

Aylett, M. P., & Turk A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech*, 47:1, 31-56.

Barabási, A. L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439), 509–512.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1), 92–111.

Bonatti, L. L., Pena, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, 16(6), 451–459.

Brown, A. (1988). Functional load and the teaching of pronunciation. *Tesol Quarterly*, 22(4), 593-606.

Bybee, J. (2003). *Phonology and Language Use*. Cambridge University Press.

Canepari, L. (2009). *Dizionario di pronuncia italiana*. Bologna: Zanichelli Editore.

Caramazza, A., Chialant, D., Capasso, R., & Miceli, G. (2000). Separable processing of consonants and vowels. *Nature*, 403(6768), 428-430.

Carnevali, S. (2009). Fonetica, downloaded from http://www.webalice.it/sandro.carnevali2011/

Carter, D. M. (1987). An information-theoretic analysis of phonetic dictionary access. *Computer Speech & Language*, 2(1), 1-11.

CC-CEDICT Dictionary, downloaded on 30 Nov 2012 from http://cc-cedict.org/wiki/.

Cercle Linguistique de Prague, (1931). *Réunion phonologique internationale : 18-21 / XII 1930, tenue à Prague. Travaux du Cercle linguistique de Prague*, 4. Prague: Jednota ceskoslovenskych matematiku a fysiku.

Cholin, J., Levelt, W. J. M., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition*, 99(2), 205–235.

Crothers, J. (1978). Typology and universals of vowel systems in phonology. In Greenberg, J. H., Ferguson, C. A., & Moravcsik, E. A. (Eds.) *Universals of human language, vol. 2*. Stanford, CA: Stanford University Press, 93-152.

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & Van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition*, 28(5), 746–755.

Da Tos, M. (2013). The Italian FINIRE-type verbs: a case of morphomic attraction. *The Boundaries of Pure Morphology: Diachronic and Synchronic Perspectives*, *4*, 45.

Delle Luche, C., Poltrock, S., Goslin, J., New, B., Floccia, C., & Nazzi, T. (2014). Differential processing of consonants and vowels in the auditory modality: A cross-linguistic study. *Journal of Memory and Language*, 72, 1–15.

Easterday, S., Timm, J., & Maddieson, I. (2011). The effects of phonological structure on the acoustic correlates of rhythm. *ICPhS XVII*, 623-626.

Ferrer i Cancho, R., & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment.*

Fogerty, D., & Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, 131(2), 1490–1501.

Fry D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, 5, 171-189.

Gelas, H., Besacier, L., & Pellegrino, F. (2012). Developments of Swahili resources for an automatic speech recognition system. In *Proceedings of the Third International Workshop on Spoken Languages Technologies for Under-resourced Languages*, 94-101.

Gerlach, M., & Altmann, E. G. (2013). Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, 3(2), 021006.

Gess, R., Lyche, C., & Meisenburg, T. (Eds.). (2012). *Phonological Variation in French: Illustrations from three continents*. John Benjamins Publishing.

Goldsmith, J. (2000). On Information theory, entropy, and phonology in the 20th century. *Folia Linguistica*, 34(1-2), 85-100

Hall, D. C. (2011). Phonological contrast and its phonetic enhancement: dispersedness without dispersion. *Phonology*, 28(01), 1–54.

Havy, M., & Nazzi, T. (2009). Better processing of consonantal over vocalic information in word learning at 16 months of age. *Infancy*, 14(4), 439–456.

Hockett, C. F. (1955). *A manual of phonology*, Waverly Press: Baltimore.

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

Hombert, J. M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, 37-58.

Hua, Z., & Dodd, B. (2000). The phonological acquisition of Putonghua (modern standard Chinese). *Journal of Child Language*, 27(01), 3-42.

Hyman, L. M. (2008). Universals in phonology. *The Linguistic Review*, 25, 83-137.

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

Jäger, G. (2012). Power laws and other heavy-tailed distributions in linguistic typology. *Advances in Complex Systems*, 15(03n04).

Jakobson, R. (1931). Principes de phonologie historique, In Troubetzkoy, N.S. *Principes de phonologie*. Paris, Klincksieck, 1976, 315-336.

Jakobson, R. (1941; 1962). Kindersprache, Aphasie und allgemeine Lautgesetze. Reprinted in *Selected Writings I*. Mouton, The Hague, 328-401.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4), 824–843.

Johnson, K. (1996). Speech perception without speaker normalization. In Johnson, K. and Mullennix (Eds.) *Talker Variability in Speech Processing*. San Diego. Academic Press.

Kello, C. T., & Beltz, B.C. (2009). Scale-free networks in phonological and orthographic wordform lexicons. In *Approaches to Phonological Complexity*, Pellegrino, F., Marsico, E., Chitoran, I. & Coupé, C. (Eds.), Phonology & Phonetics Series vol. 16, Berlin, New York, Mouton de Gruyter, 171-190.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in cognitive sciences*, 14(5), 223-232.

Kewley-Port D, Burkle TZ, Lee JH (2007) Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing impaired listeners. *The Journal of the Acoustical Society of America*, 122, 2365–2375.

Kim, S., Yi, H., Yu, C. and Han'guk Pangsong Kongsa. (1993). Py'ojun Han'gugŏ parŭm taesajŏn =: A Korean pronunciation dictionary, Sŏul T'ŭkpyŏlsi: Ŏmun'gak.

King, R. D. (1967). Functional Load and Sound Change, *Language*, 43:4, 831-852.

Kissling, E. M. (2012). Cross-linguistic differences in the immediate serial recall of consonants versus vowels. *Applied Psycholinguistics*, 33(03), 605–621.

Kronrod, Y., Coppess, E., & Feldman, N. H. (2012). A unified model of categorical effects in consonant and vowel perception. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 629-634.

Kučera, H. (1963). Entropy, redundancy and functional load in Russian and Czech. *American Contributions to the Fifth International Congress of Slavists*, Mouton & Company: The Hague, 191-219.

Ladefoged, P. (2001). *Vowels and consonants: An introduction to the sounds of languages*. Oxford: Blackwells.

Ladefoged, P., & Maddieson, I. (1996) *The sounds of the world's languages*. Blackwells, Cambridge.

Labov, W. (2001). *Principles of linguistic change*, Vol. II: Social factors. Oxford: Blackwell.

Learner, Z. (2013). JyutDict, downloaded on 3 Mars 2013 from http://zhongwenlearner.com/down-loads/jyutdict/.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(01), 1–38.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In Schölkopf, B., Platt, J. & Hoffman, T. (Eds.) *Advances in Neural Information Processing System.* Cambridge, MA: MIT Press, 849-856.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1975). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-368.

Liljencrants, J. & Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language*, 48, 839-862.

Lindblom, B. (1986). Phonetic universals in vowel systems. In Ohala, J. J. & Jaeger, J. J. (Eds.), *Experimental phonology*. Orlando, FL: Academic Press, 13.

Lindblom, B. & Maddieson, I. (1988). Phonetic universals in consonant systems. In Hyman, L. M. & Li, C. N. (Eds.) *Language, speech and mind*, London: Routledge, 62-78.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, 19(1), 1.

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.

Maddieson, I. (1984). *Patterns of sounds.* Cambridge, MA: Cambridge University Press.

Maddieson, I. (2007). Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. In M-J. Solé, P. Beddor, & M. Ohala, (Eds.), *Experimental Approaches to Phonology*. Oxford University Press, Oxford and New York: 93-103.

Maddieson, I. (2013). Syllable Structure. In: Dryer, Matthew S. & Haspelmath, Martin (Eds.) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/12, Accessed on 2014-10-21)

Maddieson, I., & Precoda, K. (1990). Updating UPSID. *Working Papers in Phonetics 74*, Department of Linguistics, UCLA, UC Los Angeles

Maddieson, I., Flavier, S., Marsico, E., Coupé, C., & Pellegrino, F. (2013). LAPSyD: Lyon-Albuquerque phonological systems database. In *Proceedings of Interspeech 2013*. 3022-3026.

Marsico, E., Maddieson, I., Coupé, C., & Pellegrino, F. (2003). Investigating the "hidden" structure of phonological systems. In *Proceedings of the 30th Annual Meeting of the Berkeley Linguistics Society*, 256-267.

Martindale, C., Gusein-Zade, S. M., McKenzie, D., & Borodovsky, M. Y. (1996). Comparison of equations describing the ranked frequency distributions of graphemes and phonemes. *Journal of Quantitative Linguistics*, 3(2), 106–112.

Martinet, A. (1938). La phonologie. *Le français moderne*, 6, 131-146.

Martinet, A. (1955). *Économie des changements phonétiques. Traité de phonologie diachronique*, Francke: Berne.

Max Planck Institute for Psycholinguistics, *WebCelex*, retrieved on 18 Mars 2013 and on 6 August 2014 from http://celex.mpi.nl.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174), 1006-1010.

Morales, F. & Taylor, J. R. (2007). *Learning from relative frequency*. Available as LAUD (Linguistic Agency, University of Duisburg) preprint, paper no. 690.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34(4), 520-531.

National Institute for Japanese Language and Linguistics and National Institute of Information and Communications Technology, *The corpus of spontaneous Japanese (CSJ)*, Third printing, 2011.

Nazzi, T. (2005). Use of phonetic specificity during the acquisition of new words: Differences between consonants and vowels. *Cognition*, 98, 13-30.

Nazzi, T., & New, B. (2007). Beyond stop consonants: Consonantal specificity in early lexical acquisition. *Cognitive Development*, 22(2), 271–279.

Nazzi, T., Floccia, C., Moquet, B., & Butler, J. (2009). Bias for consonantal information over vocalic information in 30-month-olds: Cross-linguistic evidence from French and English. *Journal of Experimental Child Psychology*, 102(4), 522–537.

Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e linguaggio*, 2(2), 203-230.

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. http://www.lexique.org.

New, B., Araújo, V., & Nazzi, T. (2008). Differential processing of consonants and vowels in lexical access through reading. *Psychological Science*, 19(12), 1223–1227.

NJStar Software Corp (2013). Chinese Word Processor v.5.30, downloaded from http://www.njstar.com/cms/njstar-chinese-word-processor/.

Obleser, J., Leaver, A., VanMeter, J., & Rauschecker, J. P. (2010). Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. *Auditory Cognitive Neuroscience*, 1, 232.

Oh, Y. M., Pellegrino, F., Coupé, C., & Marsico, E. (2013). Cross-language comparison of functional load for vowels, consonants, and tones. In *Proceedings of Interspeech 2013*, Lyon, France, 3032-3036.

Owren MJ, Cardillo GC (2006) The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *The Journal of the Acoustical Society of America*, 119, 1727–1739.

Piantadosi, S. T., Tily, H. J. and Gibson, E. (2009). The communicative lexicon hypothesis. *Proceedings of the 31st annual meeting of the Cognitive Science Society* (CogSci09), 2582-2587.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. & Hopper, P. (Eds.) *Frequency Effects and Emergent Grammar*. Amsterdam: John Benjamins Publishing, 137-157.

Pierrehumbert, J. (2003). Probabilistic phonology: Discrimination and robustness. In Bod, R., Hay, J., & Jannedy, S. (Eds.) *Probability Theory in Linguistics*. The MIT Press. Cambridge, MA, 177-228.

Pye, C., Ingram, D., & List, H. (1987). A comparison of initial consonant acquisition in English and Quiché. In Nelson, K. & Van Kleek, A. (Eds.), *Children's Language*, 6, 175-190. Hillsdale: Erlbaum.

Research Centre on Linguistics and Language Information Sciences, The Hong Kong Institute of Education (2013). *A linguistic corpus of mid-20th century Hong Kong Cantonese*. Retrieved on 1 Mars 2013 from http://hkcc.livac.org.

Rose, Y. (2009). Internal and external influences on child language productions. In Pellegrino, F., Marsico, E., Chitoran, I., & Coupé, C. (Eds.), *Approaches to phonological complexity*. Berlin: Mouton de Gruyter, 329-351.

Scharinger, M., Idsardi, W. J., & Poe, S. (2011). A comprehensive three-dimensional cortical map of vowel space. *Journal of cognitive neuroscience*, 23(12), 3972-3982.

Schilling, H. H., Rayner, K., & Chumbley, J. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, 26(6), 1270–1281.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIGDAT-Workshop.* Dublin, Ireland.

Schwartz, J. L., Boë, L. J., Vallée, N., & Abry, C. (1997). Major trends in vowel system inventories. *Journal of Phonetics*, 25(3), 233–253.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-423, 623-656.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S. (Eds.) *WaCky! Working papers on the web as corpus*, Gedit, Bologna, http://corpus.leeds.ac.uk/query-zh.html.

Sheik, A. (2013). CantoDict, http://www.cantonese.sheik.co.uk/.

Steels, L., & McIntyre, A. (1998). Spatially distributed naming games. *Advances in complex systems*, 1(04), 301-323

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4), 1872–1891.

Stilp, C. E., & Kluender, K. R. (2010). Cochlea-scaled spectral entropy, not consonants, vowels, or time, best predicts speech intelligibility, *Proceedings of the National Academy of Sciences*, 107, 12387–12392.

Stokes, S. and Surendran, D. (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. & Niyogi, P. (2003). Measuring the usefulness (functional load) of phonological contrasts. *Technical Report TR-2003-12*, Department of computer science, University of Chicago.

Surendran, D. & Levow, G. A. (2004). The functional load of tone in Mandarin is as high as that of vowels. In *Proceedings of Speech Prosody 2004*, Japan.

Surendran, D. & Niyogi, P. (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In Thomsen, O. N., (Ed.), *Competing Models of Linguistic Change: Evolution and Beyond*. Amsterdam and Philadelphia: John Benjamins, 43-58.

Tanaka-Ishii, K. (2012). Information Bias Inside English Words. *Journal of Quantitative Linguistics*, 19(1), 77-94.

Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding words and rules in a speech stream functional differences between vowels and consonants. *Psychological Science*, 19(2), 137–144.

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Göttingen: Vandenhoek and Ruprecht.

Universität Leipzig, Leipzig corpora collection (LCC), http://corpora.informatik.uni-leipzig.de.

Vallée, N. (1994). *Systèmes vocaliques : de la typologie aux prédictions*. PhD dissertation, Université Stendhal, Grenoble, France.

Van Severen, L., Gillis, J. J., Molemans, I., Van Den Berg, R., De Maeyer, S., & Gillis, S. (2012). The relation between order of acquisition, segmental frequency and function: the case of word-initial consonants in Dutch. *Journal of child language*, 40(04), 703-740.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research: JSLHR*, 51(2), 408–422.

Vitevitch, M.S., Chan, K.Y., Goldstein, R. (2014). Insights into failed lexical retrieval from network science. *Cognitive Psychology*, 68, 1-32.

Walsh, M., Möbius, B., Wade, T., & Schütze, H. (2010). Multilevel exemplar theory. *Cognitive Science*, 34(4), 537–582.

Wang, W. Y. (1967). The measurement of functional load. *Phonetica,* 16(1), 36-54.

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

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, Mass.: Addison-Wesley Press.

APPENDIX 1

Phonemic inventories of nine languages (obtained from each corpus analyzed and may contain some phonemes from the transcription of loanwords)

Language	CMN	DEU	ENG	FRA	ITA	JPN	KOR	SWH	YUE
V	i	iː	iː	i	i	i	i	i	i
	y	yː	uː	y	u	iː	ɯ	u	iː
	u	uː	I	u	e	ɯ	u	e	y
	ə	ɪ	ʊ	e	ø	ɯː	e	o	yː
	o	ʏ	ə	ø	o	e	o	a	u
	ɚ	eː	ɜː	o	ɛ	eː	ɛ		uː
	a	øː	ɛ	ə	ɔ	o	ʌ		e
		ʊ	ʌ	ɛ	a	oː	a		o
		oː	ɔː	œ		a			ɛː
		ə	æ	ɔ		aː			œː
		ɜː	æ̃	ɛ̃					ɔː
		ɛ	æ̃ː	œ̃					ɐ
		ɛː	ɒ	ɔ̃					aː
		œ	ɑː	a					
		œ̃ː	ɑ̃ː	ɑ̃					
		ʌ	ɒ̃ː						
		ɔː	eI						
		ɔ	aI						
		æ	ɔI						
		æ̃	əʊ						
		æ̃ː	aʊ						
		a	Iə						
		aə	ɛə						
		ɒ̃ː	ʊə						
		ã:							
		eɪ							
		aɪ							

		ɔɪ aʊ ai au ʊy							
Language	CMN	DEU	ENG	FRA	ITA	JPN	KOR	SWH	YUE
C	p t k pʰ tʰ kʰ ts tʂ tɕ tsʰ tʂʰ tɕʰ f s ʂ ʐ ɕ x w ɥ j l m n ŋ	p t k b d g p͡f t͡s v t͡ʃ d͡ʒ m ŋ f v s z ʃ ʒ X,ç h l R,r w j	p t k b d g f v θ ð s z ʃ ʒ x, ç h t͡ʃ d͡ʒ m n ŋ l r, R w j	p t k b d g f v s z ʃ ʒ ʁ m n ɲ ŋ l R w ɥ j	p t k b d g t͡s d͡z t͡ʃ d͡ʒ f v θ s z ʃ ʒ m n ɲ l r ʎ w j	p t k b d g s z h m n r w j	p t c k pʰ tʰ cʰ kʰ b d g d͡z m n ŋ s sʰ h l w ɯɥ j	p t c k b d ɟ g m n mv nd ɲɟ ŋg mb nz f v θ ð s z ʃ x ɣ h l r w j	p t k kʷ pʰ tʰ kʰ kʷʰ t͡s t͡sʰ f s h m n ŋ l w j

APPENDIX 2

Below is a toy example that illustrates the differences between the configurations INF/TOKEN, INF/TYPE, LEM/TOKEN and LEM/TYPE.

The starting point is a fictitious corpus based on an extraction of entries of the WebCelex English corpus:

Inflected form	**Lemma**	**Phonetic form**	**Grammatical category**	**Frequency**
beautiful	beautiful	'bju:-tə-fʊ̈l	Adjective	2075
beautifully	beautifully	'bju:-tə-flI	Adverb	278
drink	drink	'drIŋk	Verb	728
drinks	drink	'drIŋks	Verb	111
drink	drink	'drIŋk	Noun	1414
drinks	drink	'drIŋks	Noun	440
drinker	drinker	'drIŋ-kəR	Noun	30
drinkers	drinker	'drIŋ-kəRs	Noun	44
drank	drink	'dræŋk	Verb	620

For each corpus, entries are merged on the basis of similar phonetic forms, regardless of grammatical categories. For a set of entries with an identical phonetic form, the frequency of the resulting entry is equal to the sum of the frequencies of the merged entries.

To build the INF/TOKEN corpus, one therefore only needs to merge identical phonetic forms, more precisely here i) /'drIŋk/ as a verb and as a noun, ii) /'drIŋks/ as a verb and as a noun:

Inflected form	**Phonetic form**	**Frequency**
beautiful	'bju:-tə-fʊ̈l	2075
beautifully	'bju:-tə-flI	278
drink	'drIŋk	2142 (728+1414)
drinks	'drIŋks	551 (111+440)
drinker	'drIŋ-kəR	30
drinkers	'drIŋ-kəRs	44
drank	'dræŋk	620

The INF/TOKEN corpus

To obtain the LEM/TOKEN corpus, we first merge the entries of the initial set according to their lemmas. The frequency of a lemma form is equal to the sum of the frequencies of the corresponding inflected forms:

Lemma	**Phonetic form**	**Grammatical category**	**Frequency**

beautiful	'bju:-tə-fʊ̈l	Adjective	2075
beautifully	'bju:-tə-flI	Adverb	278
drink	'drIŋk	Verb	1459 (728+111+620)
drink	'drIŋk	Noun	1854 (1414+440)
drinker	'drIŋ-kəR	Noun	74 (30+44)

Intermediate corpus while building the LEM/TOKEN corpus

The second step is to merge entries according to their phonetic forms, as done previously for the INF/TOKEN corpus:

Lemma	**Phonetic form**	**Frequency**
beautiful	'bju:-tə-fʊ̈l	2075
beautifully	'bju:-tə-flI	278
drink	'drIŋk	3313 (1459+1854)
drinker	'drIŋ-kəR	74 (30+44)

The LEM/TOKEN corpus

Considering types rather than tokens amounts to equating all frequencies to 1. We can therefore easily derive the INF/TYPE corpus from the previous INF/TOKEN corpus. Note that equating the frequencies to 1 should take place *after* extracting the 20,000 most frequent entries, as mentioned in section 2.3 (this is not relevant for our small toy corpus). The LEM/TYPE corpus is obtained from the LEM/TOKEN corpus the way that the INF/TYPE corpus is derived from the INF/TOKEN corpus:

Inflected form	**Phonetic form**	**Frequency**
beautiful	'bju:-tə-fʊ̈l	1
beautifully	'bju:-tə-flI	1
drink	'drIŋk	1
drinks	'drIŋks	1
drinker	'drIŋ-kəR	1
drinkers	'drIŋ-kəRs	1
drank	'dræŋk	1

The INF/TYPE corpus

APPENDIX 3

List of the contrasting pairs of vowels/consonants (ranked by increasing FL) for the nine languages under study, used in the simulation presented in Section 4.1 to estimate the relative loss of entropy when gradually coalescing lower-FL segments with higher-FL segments.

Language	CMN	DEU	ENG	FRA	ITA	JPN	KOR	SWH	YUE
V	o→a	æː→iː	ɒ̃ː→ɒ	ə→a	ø→a	aː→ɯ	ɯ→a	e→a	y→u
	ɘ→y	ɒ̃ː→uː	ɑ̃ː→ɒ	œ→ɛ	ɛ→a	ɯː→i	ɛ→a	o→a	i→u
	y→i	ãː→Y	ʊə→ɔː	ɔ→a	u→o	iː→ɪ	u→a	u→i	œː→e
	i→u	ə→ɪ	ə→I	ɛ̃→a	ɔ→a	eː→a	ʌ→i	i→a	uː→aː
	u→ə	øː→oː	ɔI→eI	œ̃→ɛ	o→a	ɯ→a	e→i		yː→iː
	a→ə	œ→ɔ	Iə→ɔː	y→ɛ	i→e	oː→a	o→i		u→ɐ
		ɛː→aː	ʊ→æ	o→a	e→a	i→a	a→i		e→o
		ʊy→au	ɛə→ɔː	u→a		o→a			o→ɐ
		Y→ɛ	ɜː→ɔː	ɔ̃→ɛ		e→a			ɛː→ɔː
		ʊ→a	aʊ→eI	i→ɛ					iː→aː
		yː→aː	ɑː→eI	ɛ→e					ɐ→aː
		uː→aː	uː→iː	ɑ̃→e					aː→ɔː
		au→ai	ʌ→æ	ø→e					
		ɔ→ɛ	ɒ→I	a→e					
		oː→iː	ɔː→I						
		eː→aː	ɛ→I						
		aː→iː	əʊ→eI						
		ɛ→a	æ→I						
		ɪ→a	I→eI						
		ai→a	iː→eI						
		iː→a	aI→eI						

Language	CMN	DEU	ENG	FRA	ITA	JPN	KOR	SWH	YUE
C	ŋ→n	ʒ→b	x, ç→p	ŋ→t	θ→n	f→k	tʰ→n	x→h	kʷʰ→t͡sʰ
	ɥ→j	t͡ʃ→p	ʒ→s	ɥ→ʁ	ʒ→f	p→k	kʰ→g	mv→v	kʰ→l
	f→ʂ	d͡ʒ→b	ŋ→d	ɲ→ʁ	d͡z→b	b→k	c→b	θ→ɟ	pʰ→k
	pʰ→m	p͡f→t	θ→d	w→ʁ	w→r	c→k	p→d	ɣ→w	ŋ→k
	ʐ→ʂ	ŋ→s	ʤ→t	z→ʁ	ɲ→b	z→k	w→j	ð→k	kʷ→t͡s
	tɕʰ→tɕ	j→v	t͡ʃ→k	g→t	j→r	r→k	s→d	nz→c	f→k
	kʰ→ʂ	ʃ→k	j→l	j→ʁ	ʃ→v	j→n	ŋ→n	ɲɟ→t	w→m

tsʰ→l	p→n	g→k	ʃ→ʁ	t͡s→t	h→k	pʰ→g	g→t	tʰ→s
s→ʂ	t͡s→t	v→d	b→k	g→d	w→g	k→d	f→t	n→m
w→m	g→b	p→k	f→v	z→n	d→k	ɯ→g	ʃ→k	p→t
ɕ→tɕ	h→v	r, R→z	ʁ→l	b→t	g→k	j→sʰ	r→t	l→t
tʂʰ→ʂ	b→R,r	ʃ→t	ʒ→s	d͡ʒ→t	m→n	cʰ→sʰ	d→k	j→s
tɕ→ʂ	k→n	k→t	v→t	p→m	n→t	d͡ʑ→sʰ	nd→k	t͡sʰ→s
ts→ʂ	l→R,r	f→b	k→p	f→m	t→k	t→b	v→k	m→t
x→t	X,ç→s	l→t	p→t	t͡ʃ→s	s→k	h→sʰ	ŋg→t	t→k
m→l	f→s	b→t	t→s	r→t		b→g	mb→n	h→k
j→ʂ	t→n	h→w	n→s	v→t		m→g	b→k	s→ts
k→tʂ	s→n	z→d	m→s	ʎ→d		d→g	ɻ→w	k→ts
tʰ→l	v→z	w→t	d→s	m→k		sʰ→g	p→k	
n→l	z→R,r	d→t	l→s	t→n		l→n	h→k	
p→t	d→R,r	s→t		k→s		g→n	s→k	
tʂ→ʂ	m→R,r	ð→m		s→n			t→n	
ʂ→l	R,r→n	m→t		n→d			m→k	
l→t		n→t		l→d			k→l	
							c→w	
							z→l	
							l→n	
							w→j	
							j→n	