Cross-language Comparison of Functional Load for Vowels, Consonants, and Tones

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

Laboratoire Dynamique du Langage, Université de Lyon and CNRS, France

{yoon-mi.oh; francois.pellegrino}@univ-lyon2.fr,

{christophe.coupe; egidio.marsico}@ish-lyon.cnrs.fr

Abstract

The notion of functional load (FL) quantifies the role a phonological contrast plays in keeping words distinct in a given language. Several studies have emphasized its potential impact on language evolution and acquisition, and FL has repeatedly been mentioned as a useful tool to supplement phonological descriptions for more than seventy years. It is nevertheless still rarely explored and this paper is a contribution to filling this gap.

By adopting an information-theory approach and a measure of FL proposed by Hockett (1955), we performed a corpus-based comparison of three non-tonal (English, Japanese, Korean) and two tonal languages (Cantonese and Mandarin). We calculated FLs carried by segmental (vowels and consonants) contrasts and tonal contrasts (in Cantonese and Mandarin). We also evaluated the total FL associated with the vocalic system as a whole, the consonantal system as a whole, and the tonal system (when applicable).

Our results suggest that i) the distributions of FLs in a phonological system are very uneven, with only a few prominent contrasts, and ii) the existence of a tonal system does not reduce the importance of vowel and consonantal contrasts, even though tone contrasts are as important as vowel contrasts in Cantonese and Mandarin.

Index Terms: Cantonese, Cross-language Study, English, Functional Load, Japanese, Korean, Mandarin, Tones, Phonological System

1. Introduction

Phonological contrast and opposition were central concepts within the Prague School [1], along with the idea that the importance of each specific contrast in a phonological system may differ from one language to another. Such a notion, called Functional Load (FL, henceforth), was further developed by Martinet, who also suggested that FL may play a role in language evolution [2], [3]. According to his hypothesis, phonemes involved in high-FL contrasts would be less prone to change than those involved in low-FL contrasts. Hockett also considered that "The function of a phonemic system is to keep the utterances of a language apart" and observed that "Some contrasts between the phonemes in a system apparently do more of this job than others" [4].

Since then, a few studies have hypothesized or assessed the role of FL in various areas: linguistic typology, description of phonological systems, automatic speech recognition, child language and second-language acquisition, sound change from diachronic and synchronic perspectives, identification of articulatory and perceptual constraints on phonological systems, etc. ([5-9] among others). However, the role of FL is still debated, since several diachronic studies do not support Martinet's hypothesis (e.g. [5], [6]).

Following Hockett [4], we consider that FL may be especially useful to shed light on the organization of phonological systems and on the relative weight associated to their components. In this paper, our aim is therefore to evaluate the usefulness of a quantitative approach to FL as a tool to describe and compare phonological systems.

Five languages (Cantonese, English, Japanese, Korean and Mandarin) were compared and analyzed through a quantitative corpus-based approach. These languages were chosen in order to provide some variations in phonology (tonal vs. non-tonal languages) to answer two research questions:

- Are the FL carried by segmental components (vowels and consonants) comparable among languages?
- What is the FL associated with tonal systems?

Corpora and methods are described in the next section while several results are presented and discussed in Section 3.

2. Data and Method

2.1. Corpus description and preprocessing

Each corpus was collected separately for the five languages (Cantonese, English, Japanese, Korean and Mandarin), as detailed in Table 1.

Language	ISO 639-3 Code	Source	Corpus Size (#Tokens)	Phonological System	
Cantonese	yue	[10]	144.1k	V C T	10 19 6
English	eng	[11]	18.6M	V C	14 24
Japanese	jpn	[12]	2.4M	V C	10 17
Korean	kor	[13]	2.4M	V C	8 22
Mandarin	cmn	[14]	281.7M	V C T	8 22 5

Table 1. *Corpus description. For each language the size of its phonological system is provided (V: #vowels, incl. diphthongs for English; C: #consonants; T: #tones, if applicable)*

After cleaning erroneous entries (e.g. trivial errors due to automatic transcription), several preprocessing steps were necessary, depending on each language.

For Mandarin and Cantonese, we relied on public domain dictionaries and software to get the pinyin and jyutping transcriptions respectively. For Mandarin, the CC-CEDICT dictionary was used [15]; additionally, when an entry of the corpus was missing in it, we used NJStar Chinese Word Processor to get the transcription [16]. For Cantonese, we compared the transcriptions provided by CantoDict [17] and JyutDict [18] to choose an appropriate transcription. When

differences between dictionaries reflected on-going changes (like the deletion of initial /ŋ/), the most traditional pronunciations were kept. We discarded entries of the corpus for which no corresponding entry was available in the dictionaries, which reduced the size of the wordlist from 8,541 to 5,713. Once pinyin and jyutping transcriptions were obtained, Dr. F. Wang's assistance helped us to get the phonological transcriptions of the various possible syllables. The Japanese corpus was originally transcribed in katakana by native speakers. We converted it into a phonological transcription thanks to a list of phonemic entities which correspond with morae in katakana [12].

In the case of Korean, the corpus did not contain any transcription. It was consequently transcribed by the first author, adopting the Revised Romanization of Korean, and then converted into IPA by consulting a Korean pronunciation dictionary [19]. The data of English consist of a large text corpus which was transcribed by using an automatic grapheme-to-phoneme conversion [11].

2.2. Methods

2.2.1. Definition of Functional Load

Following Hockett [4], a language *L* was considered as a source of sequences made of words *w* taken from a finite set of size N_L. The amount of information of language *L* was estimated in terms of Shannon entropy *H(L)* [20]:

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

Where $p(w_i)$ is the probability of word w_i, estimated from a large corpus.

Following Surendran & Niyogi [6], we implemented the definition of FL given by Carter [21] and derived from Hockett's initial proposal [4]. The FL of a contrast *x*/*y*, *FL*(*x*,*y*), was defined as the relative difference (in percentage) in entropy between two states of language *L*: the observed state *L* and a fictive state L^*_{xy} in which the contrast is neutralized (or coalesced, in Hockett's terminology). FL therefore quantifies the perturbation induced by merging *x* and *y*, in terms of increase of homophony and of changes in the distribution of word frequencies:

$$FL(x,y) = \frac{H(L) - H(L^*_{xy})}{H(L)} \quad (2)$$

FL(x,y) is hence defined at the level of phonemic *contrasts*. In addition, one can also focus on the level of the *phonemes* themselves, by summing *FL(x,y)* over all the contrasts in which a phoneme *x* is involved:

$$FL'(x) = \frac{1}{2}\sum_{y} FL(x,y) \quad (3)$$

With the normalization factor ½ applied to ensure that:

$$\sum_{x} FL'(x) = \sum_{x,y \neq x} FL(x,y) \quad (4)$$

2.2.2. Corpus Analysis

For each language, the FL was independently computed for each consonantal contrast, each vowel contrast (incl. diphthongs for English), and each tonal contrast for Cantonese and Mandarin. We furthermore computed the FL carried by natural subsystems.

More precisely, the FL associated with the vocalic system as a whole (FL_V) was calculated by defining a fictive language *L** in which all vowel segments were coalesced. For instance, this procedure applied to an English corpus results in merging the 3 distinct words *hat*, *hit*, and *hut* into a common form /h*t/, where '*' denotes the coalesced segment. This approach was also respectively applied to the consonantal system (FL_C) and the tonal system (FL_T). Table 2 illustrates this procedure for some Korean and Mandarin data.

	Korean	Mandarin
Sample Data	발 /bal/ 팔 /pʰal/ 반 /ban/ 번 /bʌn/	半 /pan4/ 判 /pʰan4/ 盘 /pʰan2/ 棒 /paŋ4/ 蹦 /peŋ4/
***FL*(x,y)**	/ba*/ /pʰa*/ /bʌ*/	/pa*4/ /pʰa*4/ /pe*4/ /pʰa*2/
***FL*$_C$**	/*a*/ /*ʌ*/	/*a*4/ /*e*4/ /*a*2/
***FL*$_V$**	/b*l/ /pʰ*l/ /b*n/	/p*n4/ /pʰ*n4/ /p*ŋ4/ /pʰ*n2/
***FL*$_T$**	n.a.	/pan*/ /pʰan*/ /paŋ*/ /peŋ*/

Table 2. *Example of Functional Load computation. The coalescence process is illustrated for one consonantal contrast (/n,l/ in Korean and /n,ŋ/ in Mandarin), and for the consonantal, vocalic, and tonal systems.*

Several factors may influence the estimation of the language entropy and, consequently, the computation of FLs. The corpus size may especially affect word frequencies estimation, and taking very low-frequency words and hapaxes into account may also influence *H(L)*. In the next section, results are reported considering the 20,000 most frequent phonological words in each language, except for Cantonese for which we were limited to the 5,000 most frequent words due to the relatively small size of the corpus (see Table 1).

For each language, syllabic boundaries were 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/ resulted in two distinct entries /*æ.**ɪ*/ and /*æ*.*ɪ*/, while they would merge into a single entry /*æ**ɪ*/ if syllable boundaries were not considered.

3. Results

During a preliminary analysis (§3.1), we evaluated the relationship between the FLs observed for the segments and their relative frequencies in the corpus. In the second subsection (§ 3.2), we then compared the inner organization of phonological systems across languages, in terms of individual phonemic and tonal contrasts. Finally, we conducted a comparison of the relative weights associated to the three classes of vowels, consonants and tones (§3.3).

3.1. Relationship between FL and Frequency

A reasonable assumption is that frequent phonemes in a language will exhibit a higher FL than less frequent ones, since they are more utilized in words. The Pearson correlation between FL'(x) and the frequency of occurrence of x was calculated for each language, independently for consonant and vowel systems (Table 3). Results illustrate that this relationship is rather strong in most cases (up to $r^2 = 0.95$ for Mandarin vowels), but also that it is not straightforward and may be very limited. In Mandarin, for instance, the correlation between FL and frequency for consonants is very low, as illustrated in Figure 1. This figure shows that even though the

(linear) relationship between FL and frequency holds for most consonants, /ŋ/ and /n/ depart from it, with a much lower FL than their frequencies of occurrence would predict. Since they are the only consonants which occur as syllable codas in Mandarin, it suggests that most words differing in codas also differ in other segments or tones.

	yue	cmn	eng	jpn	kor
Consonants	0.36	0.10	0.73	0.47	0.82
Vowels	0.41	0.95	0.50	0.77	0.30

Table 3. *Correlation between FLs and frequencies for consonants and vowels (r^2).*

All in all, although the phoneme FLs can to a good extent be predicted from their frequency, other factors (coming under syllable construction, phonotactics, morphology, etc.) underlie the distinctive function within a phonological system.

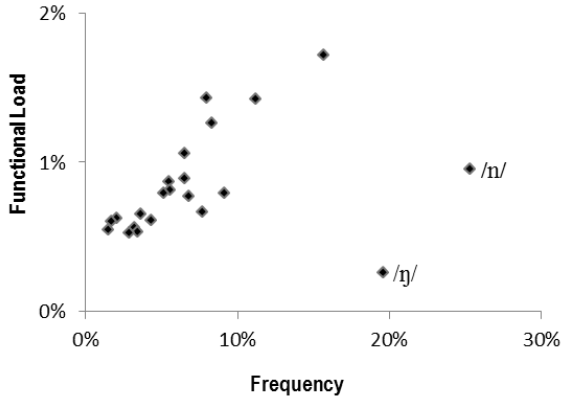

Figure 1. *Correlation between the percentage of FL and the frequency of Mandarin consonants.*

3.2. Contrasts and Segments FL

3.2.1. Contrast Distribution

FL distributions in each language were compared for the 10 vowels with highest FLs in each language (Figure 2). Despite differences in amplitude, the distributions exhibit a similar trend. Very few contrasts exhibit high FL, and the other contrasts have very low FLs. This skewed pattern is more present for Mandarin and English than for Korean.

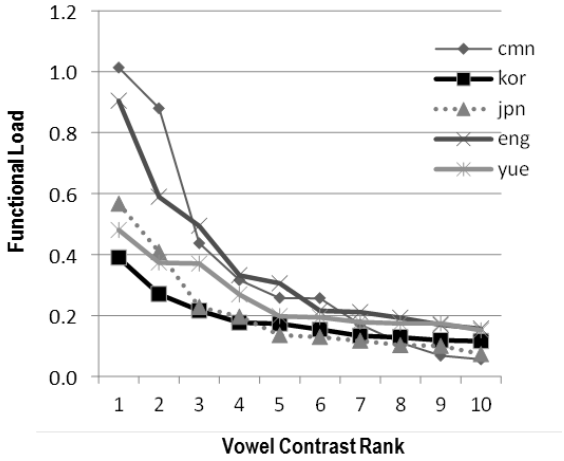

Figure 2. *Distribution of 10 vowel pairs with the highest FL*

This result is compatible with our preliminary observations made with another set of languages [9]. Results are similar for consonant contrasts (not displayed).

	yue		cmn		eng		jpn		kor	
1	ɔ:-a:	0.480	ə-a	1.013	aɪ-eɪ	0.904	e-a	0.567	i-e	0.390
2	ɛ:- ɔ:	0.373	u-i	0.880	ɪ-æ	0.589	o-a	0.409	o-i	0.270
3	o- ɐ	0.370	u-ə	0.437	eɪ-i:	0.494	i-a	0.228	i-a	0.216

Table 4. *Top vowel contrasts (ranked by decreasing FL).*

For the three major vowel contrasts (Table 4), the highest FL is associated with a contrast in terms of aperture, and is not a maximum perceptual contrast (for instance /e,a/ in Japanese rather than /i,a/). Contrasts on the front/back dimension are also present in this Table (e.g. /ɛ:,ɔ:/ in Cantonese and /u,i/ in Mandarin).

Table 5 shows the three most important consonant contrasts. Nasals, stops, affricates, fricatives, liquids and glides are present, depending on the language considered. Most of the contrasts are based on a redundant opposition in terms of phonetic features. For instance, in Japanese, /s/ and /k/ contrast both in terms of manner *and* place of articulation, in Mandarin, /t/ and /l/ differ in terms of manner *and* voicing, etc. On the contrary, a few contrasts are minimal, like /n,m/ in Cantonese or /ŋ,n/ in Mandarin, since they only differ in place. As said in §3.1, the /ŋ,n/ contrast in Mandarin is an example of a position-specific contrast since it occurs only in codas. /h,ð/ in English provides another example, with a word-initial opposition (/h/ is restricted to word-initial position, and in this position /ð/ is only encountered in grammatical words).

	yue		cmn		eng		jpn		Kor	
1	n-m	0.452	t-l	0.742	n-t	0.594	s-k	0.984	l-n	0.523
2	ts-t	0.378	ŋ-n	0.445	z-t	0.521	w-g	0.603	g-t	0.155
3	ts-k	0.346	t-ʂ	0.342	h-ð	0.430	n-t	0.504	n-g	0.143

Table 5. *Top consonant contrasts (ranked by decreasing FL).*

3.2.2. Functional load of phonological segments

Table 6 shows the five vowels with the largest FL in each language. At first glance, there is a striking diversity, both along the front/back and high/low dimensions, depending on the language. From this functional viewpoint, this selection of five vowels largely departs from a canonical system evenly distributed in the vocalic space. In other words, none of these languages heavily rests upon a maximally dispersed /i,a,u/ triplet, and each language has its specificities.

	yue		cmn		eng		jpn		kor	
1	ɔ:	0.71	u	0.86	eɪ	1.15	a	0.76	i	0.58
2	a:	0.65	i	0.85	aɪ	1.06	e	0.50	a	0.47
3	ɐ	0.65	ə	0.83	i:	1.05	o	0.48	o	0.47
4	i:	0.45	a	0.77	ɪ	0.91	i	0.33	e	0.36
5	ɛ:	0.39	y	0.27	æ	0.74	o:	0.25	ʌ	0.27

Table 6. *Highest-FL vowels (ranked by decreasing FL').*

Cantonese distinctions are mostly based on mid to open vowels, while the reverse is observed for Mandarin. English distinctions seem to favor front vowels. English, however, exhibits higher values than the other languages, and even its 5th vowel /æ/ (*FL'*(æ)=0.74) reaches almost the highest values encountered in the other languages.

Regarding consonants (Table 7), a large diversity in terms of phonetic is also present, among and within languages. A general trend is nevertheless that obstruents seem to play a more important role than sonorants since 18 of the 25 consonants in the table are obstruents. Similarly, alveolar consonants (/ts, s, t, l, n, s^h, d/) are frequent in this table. Another trend is that the top FLs for consonants are higher than the highest FLs associated with vowels in each language.

	yue		cmn		eng		jpn		kor	
1	ts	1.36	t	1.72	t	1.76	k	1.26	n	0.78
2	k	1.27	l	1.43	n	1.47	s	0.86	g	0.60
3	s	1.07	ʂ	1.42	ð	1.31	t	0.79	l	0.50
4	h	0.96	tʂ	1.26	m	1.31	n	0.74	s^h	0.43
5	t	0.94	p	1.06	s	1.24	m	0.58	d	0.41

Table 7. *Highest-FL consonants (ranked by decreasing FL').*

3.3. Weight of phonological subsystems

The previous sections showed that there is a common trend toward an uneven use of segmental contrasts and that the five languages exhibit a large diversity in the "preferred" segments and contrasts, both for vowels and consonants. In this section, we run a cross-language comparison at the more general level of the vocalic, consonantal, and tonal subsystems.

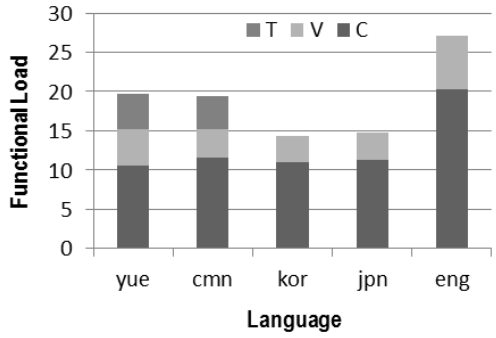

Figure 3. *Functional load carried by each phonological subsystem (Vocalic, Consonantal, and Tonal).*

Figure 3 shows the FL of the subsystems, as defined in §2.2.2. Except for English, FL_C and FL_V values are approximately similar among the languages, with no dramatic differences between tonal and non-tonal languages.

Furthermore, in the case of Mandarin, the weight of tonal contrasts FL_T is as important as that of vowels FL_V, which supports Surendran and Levow's previous result [7]. Similarly, Cantonese presents the same balance between its vocalic and tonal subsystems, even with a higher FL for tones.

Interestingly enough, English shows a quite different pattern, with much larger FLs associated to vocalic and consonantal systems than those observed with the other languages. This may be partly explained by its larger phonological system (with 38 segments in our description), but it may also translate a different weighting between this phonological level and the morphological and syntactic levels.

4. Conclusion and Perspectives

In this paper, we investigated the way five languages rely on their phonological components to distinguish among their words, with the notion of FL. Our main goal was to answer two questions: 1. Are the FL carried by segmental components (vowels and consonants) comparable among languages? 2. What is the FL associated with tonal systems?

The first result was that a large diversity was visible among the languages, and also within each phonological system. A few contrasts played a major role in each language, as shown in Figure 2 for vowels. No cross-language preference was however demonstrated in favor of maximal perceptual contrasts or specific articulations – except that obstruents seem to be more important than sonorants, and that the FL associated with consonants is higher than with vowels, as far as high-FL segments are concerned. These results are fully compatible with general trends observed in the composition of phonological systems, with more consonants than vowels, more obstruents than sonorants [22]. This result is confirmed at the level of the subsystems, with FL_C values much larger than FL_V values in the 5 languages. The group of Asian languages (Cantonese, Japanese, Korean, and Mandarin) furthermore displayed a remarkably similar pattern, despite their differences in terms of lexical prosody, morphology, and syntax (agglutinative SOV languages for Japanese and Korean vs. mainly mono- or disyllabic SVO languages for Cantonese and Mandarin). Leaving English out, one can positively answer the first question (but see below).

The answer to the second question was provided by Figure 3: FL_T is similar to FL_V for Cantonese and Mandarin, and the importance of the tonal system is not balanced by a lower role of the consonant or the vowel system.

This study exposed nevertheless that the situation is not straightforward, since English behaved quite differently, with much higher FLs associated with its consonantal and vocalic subsystems. It can reflect either methodological choices (see for instance [8] for alternative approaches) or different strategies in the way languages convey information. It definitively points toward several directions for future research. First, it underlines the need for a larger typological sample, in order to evaluate the range of variation observed in subsystem FLs in the world languages. Second, it suggests to relate the segmental and tonal levels to the syllable structure, and beyond, to the other levels of the language grammar. Several typological studies have shed light on correlations between these levels of linguistic coding (see [23] and [24] for discussions) and the approach presented in this paper can also be applied to syllable structures (by reducing the words to their syllabic structure) or to their positions in the words. In a typological perspective, this quantitative method will renew our understanding of the relative weight of each linguistic subsystem in the world languages. Finally, estimating FL with *conditional* entropy would offer a way to take context into account. It would improve the evaluation of the impact of phoneme coalescence by focusing on word pairs that remain confusable once their context of occurrence is considered. Similarly, the relevance of grammatical words in FL computation could be examined.

5. Acknowledgment

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). We also warmly thank Pr. Wang Feng (Peking University) and Pr. Andy C. Chin (The Hong Kong Institute for Education) for their precious help.

6. References

[1] Cercle Linguistique de Prague, "Thèses présentées au premier congrès de philologues slaves", Travaux du cercle linguistique de Prague 1: 5-29, Online: http://www2.unil.ch/slav/ling/textes/theses29.html, accessed on 10 Mar 2013.

[2] Martinet, A., Économie des changements phonétiques. Traité de phonologie diachronique, Francke: Berne, 1955.

[3] Martinet, A., "Some basic principles of functional linguistics", La Linguistique, 13(1): 7-14, 1977.

[4] Hockett, C. F., "The quantification of functional load: A linguistic problem", Report Number RM-5168-PR, Rand Corp., Santa Monica, 1966.

[5] King, R. D., "Functional load and sound change", Language, 43(4): 831–852, 1967.

[6] Surendran, D. and Niyogi, P., "Measuring the usefulness (functional load) of phonological contrasts", Technical report TR-2003-12, Department of computer science, University of Chicago, 2003.

[7] Surendran, D. and Levow, G. -A., "The functional load of tone in Mandarin is as high as that of vowels", In Proc. of speech prosody 2004, Nara, Japan, 99-102, 2004.

[8] Van Severen, L., Gillis, J. J., Molemans, I., van den Berg, R., De Maeyer, S., and Gillis, S., "The relation between order of acquisition, segmental frequency and function: the case of word-initial consonants in Dutch", J Child Lang, 1(1): 1-38, 2012.

[9] Pellegrino, F., Marsico, E. and Coupé, C., "La typologie des systèmes vocaliques revisitée sous l'angle de la charge fonctionnelle", Proc. of the Joint Conference JEP-TALN-RECITAL 2012, 1:JEP, Grenoble, 4-8 June, ATALA/AFCP, 617-624.

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

[11] Max Planck Institute for Psycholinguistics, WebCelex, retrieved on 18 Mars 2013 from http://celex.mpi.nl.

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

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

[14] Sharoff, S., "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, 2006.

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

[16] NJStar Chinese Word Processor v.5.30, downloaded from http://www.njstar.com/cms/njstar-chinese-word-processor/.

[17] Sheik, A., CantoDict, http://www.cantonese.sheik.co.uk/.

[18] Learner, Z., JyutDict, downloaded on 3 Mars 2013 from http://www.zhongwenlearner.com/downloads/jyutdict/.

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

[20] Shannon, C. E., "A mathematical theory of communication", Bell system technical journal, 27: 379-432 & 623-656, July and October, 1948.

[21] Carter, D. M., "An information-theoretic analysis of phonetic dictionary access", Computer Speech Lang, 2(1): 1–11, 1987.

[22] Maddieson, I. Pattern of Sounds. Cambridge, UK: Cambridge University Press, 1984.

[23] Plank, F. "The co-variation of phonology with morphology and syntax: a hopeful history". Linguistic Typology 2:2.195-230, 1998.

[24] Pellegrino F., Coupé C., and Marsico E. "Across-language perspective on speech information rate". Language 87, 539–558. 2011.