Information-Theoretic account of the epenthetic vowel in Korean

Sung-Hoon Hong

Hankuk University of Foreign Studies & Indiana University

(hongshoon@hufs.ac.kr)

Introduction

Goldsmith (2001, 2002), and Hume (2004, 2006, 2008), Hume & Bromberg (2005) proposed novel theories of markedness based on Information Theory (Shannon 1948).

- Their idea is that less marked segments or words have less information content.

Hume, Hume & Bromberg proposed Information Content (IC) (context-sensitive, more precisely) to measure the markedness of individual segments.

- The epenthetic vowel of a language has the lowest context-sensitive IC values among all vowels in that language.

Goldsmith proposed Phonological Complexity to address the wellformedness of bigger linguistic units such as words.

- The representational harmony of a word is maximized by minimizing its information complexity (Goldsmith 2002).
- Where there is morphophonemic alternation, the selection of a correct phoneme leads to complexity minimization.

The purpose of this study is to verify:

- whether the epenthetic vowel in Korean is least in information contents as argued by Hume & Bromberg,
- whether the selection of the epenthetic vowel vs. Ø in morphophonemic alternation can be characterized by complexity minimization as suggested by Goldsmith.

Information Theory

The amount of information (or information content) is measured by the unpredictable character or 'probability' that a particular element has within a given system.

The more variation and difference there is for an element, the lower the probability and the higher the information content of the element.

Essential formulas:

$$(1)\quad IC(x) = -\log_2 P(x)$$

$$(2)\quad MI(x,y) = \log_2 \frac{P(xy)}{P(x)\times P(y)}$$

How we acquired unigram/bigram probabilities

Unigram (one phoneme) and bigram (two adjacent phonemes) frequencies were obtained from *the 21st-Century Sejong Project Morpheme-Tagged Corpus*, constructed by the National Institute of the Korean Language (NIKL) from 1999 to 2004.

●This corpus was composed of 769 source files, out of which the files marked as non-contemporary or North Korean were excluded from further consideration.

●12,628,487 tokens of words (1,936,705 in types) were extracted from these files and transcribed into phonemic symbols.

- The frequencies of word forms were counted, based on which unigram and bigram frequency/probability lists were compiled.

$$(3)\quad |x| = \sum_{i=1}^{V} ([x]_{w_i} \times freq(w_i))$$

(*x* is a unigram or bigram, *w* is a word entry, *V* is the entire vocabulary, $[x]_w$ is the number of times that *x* occurs in *w*, and *freq*(*w*) is the occurrence of *w*.)

Epenthetic vowel in Information Theory

Hume & Bromberg (2005): epenthetic vowel of a language is lowest in context-sensitive IC of all vowels in that language.

The context-sensitive IC of a vowel is the sum of IC(v) and the Mutual Information (MI) of the vowel with respect to its preceding and following consonants.

$$(4)\quad IC_{cs}(v) = IC(v) + MI(pC, v) + MI(v, fC)$$

$$(5)\quad MI(C,v) = \sum_{k\in C} \log_2 \frac{P(kv)}{P(k)\times P(v)}$$

$$(6)\quad MI(v,C) = \sum_{k\in C} \log_2 \frac{P(vk)}{P(v)\times P(k)}$$

In Korean, the 'epenthetic' vowel is /ɨ/ (Sohn 1987 and Ahn 1998, among others).

- Illegal consonants and consonant clusters in L2 are adapted with /ɨ/-insertion .

  Christmas [k^{h}ɨrisɨmasɨ], trump [t^{h}ɨr«mpɨ], strike [sɨtɨraikɨ]

- Only /ɨ/ show morphophonemic alternation with Ø.

m«k-ta	m«k-ɨm	m«k-ɨni	m«k-«	'eat'
ap^{h}ɨ-ta	ap^{h}ɨ-m	ap^{h}ɨ-ni	ap^{h}-«	'sick'
cu-ta	cu-m	cu-ni	cu-«	'give'
			ki-«	'crawl'
			pe-«	'cut'

The context-sensitive IC values of Korean vowels were computed, and the result showed that /i/ has the lowest value of context-sensitive IC.

Vowel	Frequency	IC(v)	MI(*p*C, v)	MI(v, *f*C)	IC$_{CS}$(v)
ɨ	5559893	4.3423	14.7001	-21.2083	-2.1659
e	2210627	5.6729	4.6581	-4.1774	6.1536
i	5831331	4.2735	4.9605	10.6195	19.8535
o	3931431	4.8423	18.4288	-2.0648	21.2063
«	5979700	4.2372	13.8785	4.1897	22.3054
u	2674162	5.3982	16.7618	0.4044	22.5644
a	9077880	3.6350	22.0551	3.6478	29.3378
æ	1761357	6.0006	19.4430	15.1942	40.6377

Table 1. Contest-sensitive IC values of Korean vowels

Morphophonemic alternation in Information Theory

Goldsmith (2001, 2002): the representational harmony of a word is maximized by minimizing its Phonological Complexity (PC), which is the average IC value of the elements that comprise the word.

$$(7)\quad PC(x_i \cdots x_n) = \frac{1}{n}\sum_{i=1}^{n} IC(x_i) = -\frac{1}{n}\sum_{i=1}^{n} \log_2 P(x_i)$$

$$(8)\quad P(x_n \mid x_{n-1}) = \frac{|x_{n-1}x_n|}{|x_{n-1}|}$$

Regarding morphophonemic alternation, Goldsmith(2002:40) suggests that the selection of a correct phoneme leads to PC minimization.

Morphophonemic alternation to be tested: the occurrence of /ɨ/ vs. Ø in Korean verb/adjective suffixes

- There are two types: alternating and non-alternating suffixes
- After a V-final stem, suffixes occur without /ɨ/.
- After a C-final stem, alternating suffixes take /ɨ/ but non-alternating suffixes appear without /ɨ/.

	Suffixes	*C-final stems*	*V-final stems*
Alternating suffixes	-(ɨ)l«	m«k-ɨl«	cu-l«
	-(ɨ)my«n	m«k-ɨmy«n	cu-my«n
	-(ɨ)nik'a	m«k-ɨnik'a	cu-nik'a
Non-alternating suffixes	-ta	m«k-ta	cu-ta
	-ko	m«k-ko	cu-ko
	-ciman	m«k-ciman	cu-ciman

Is the selection of /ɨ/ or Ø in each form driven toward PC minimization?

- The combinations 120 verb/adjective stems (60 C-final and 60 V-final) and 6 suffixes (3 alternating, 3 non-alternating) were considered. (Selected stems were 60 most frequent stems in each category according to the frequency list compiled by NIKL (2003))
- For each suffix combination, the PC value of a form with /ɨ/ (PC1) was compared to that of a form without /ɨ/ (PC2).

The overall results were split:

- After a V-final stem with any of the suffixes considered, or after a C-final stem with an alternating suffix, the selection of /ɨ/ vs. Ø indeed conformed to PC minimization (i.e. the forms with smaller PC were selected).
- But after a C-final stem with a non-alternating suffix, PC minimization did not select correct forms consistently: a wrong form with /ɨ/ was chosen (*-ta*, *-ko*), or a correct form was selected (*-ciman*) but with so many exceptions (only 46 out of 60 made correct selections).

		V-final stems		C-final stems	
		PC1 - PC2	# of tokens PC1 > PC2	PC1 - PC2	# of tokens PC1 > PC2
Alternating	-l«	0.2089 *	56	-1.9271*	0
	-my«n	0.3586 *	60	-0.6192 *	0
	-nik'a	0.2221 *	58	-0.6883 *	0
Non-altern.	-ta	0.6454 *	60	-0.0427	23
	-ko	0.6225 *	60	-0.2080 *	8
	-ciman	0.5868 *	60	0.1184 *	46

* $p < .001$

Table 2. The comparison between the PC values of a form with /ɨ/ (PC1) and a form without /ɨ/ (PC2)

Why are non-alternating suffixes not amenable to PC minimization?

- There may be several potential reasons.
- Most non-alternating suffixes begin with an obstruent, while most alternating suffixes start with a sonorant (Kim 1989, Hong 2001). From the perspective of sonority-based syllable contact (Vennemann 1988), a syllable-initial obstruent after a consonant has a contact more stable that a syllable-initial sonorant after a consonant, which is usually adjusted by /ɨ/-insertion or assimilation in Korean.
- We have only considered 'phonemic' level. Our information about unigrams and bigrams were extracted from a phonemically transcribed corpus. An analysis based on a phonetically transcribed corpus might suggest a different result.
- The calculation of PC here is based on 'forward' conditional probability. Other methods of probability calculation might yield a different outcome.

Conclusion

By showing that Korean epenthetic vowel is indeed lowest in context-sensitive IC, this research supports Hume & Bromberg's proposal that context-sensitive IC can be a measure for segmental markedness.

Using PC for a measure of phonotactic markedness is partially successful, and a precise formula is yet to come to guarantee the selection of a right phoneme in morphophonemic alternations.

References

Ahn, Sang-Cheol (1998). An Introduction to Korean Phonology. Seoul, Korea: Hanshin Publishing.

Goldsmith, John (2001). On information theory, entropy, and phonology in the 20th century. *Folia Linguistica* XXXIV/1-2, 85-100.

Goldsmith, John (2002). Probabilistic models of grammar: phonology as information minimization. *Phonological Studies* 5, 21-46.

Hong, Sung-Hoon (2001). Richness of the Base, Lexicon Optimization, and suffix /ɨ/~ø alternation in Korean. Studies in Phonetics, Phonology, and Morphology 7, 215-242.

Hume, Elizabeth (2004). Deconstructing markedness: A predictability-based approach. Ms., Ohio State University.

Hume, Elizabeth (2006). Language specific and universal markedness: an information-theoretic approach. Paper presented at the Colloquium on Information Theory and Phonology, LSA Winter meeting 2006.

Hume, Elizabeth (2008). Markedness and the Language User. *Phonological Studies*, vol. 11.

Hume, Elizabeth and Ilana Bromberg (2005) .Predicting epenthesis: an information-theoretic account. Paper presented at the 7th Annual Meeting of the French Network of Phonology.

Kim, Young-Seok (1989). Several problems related to long vowels (written in Korean). Festschrift to Professor Lee Hey-Sook. Seoul: Hanshin Publishing.

Sohn, Hyang-Sook (1987). *Underspecification in Korean Phonology*. Ph.D. Dissertation, University of Illinois, Urbana-Champaign.

Vennemann, Theo (1988). *Preference Laws for Syllable Structure*. Berlin: Mouton de Gruyter.