# Phonotactics in Native and Sino-Korean:

## A Maximum Entropy-based phonotactic learning

Nayoung Park (Department of Linguistics, Seoul National University)

## 1. Introduction

- Phonotactics: Native speakers can judge whether certain strings are possible or not in their language.
  e.g., ***br***ick, ***bl***ick : well-formed in English vs. ***lb***ick : ill-formed
- Is the well-formedness judgment involved always categorical?
- No, it is not always the case that native speakers' intuition is all-or-nothing.
  e.g., Gradient preference in English (Berent et al. 2007)
  ***bl***if > ***bn***if > ***bd***if > ***lb***if

➢**Phonotactics in Korean nouns**

- Categorical restrictions exist. e.g., /ji, jɨ, wu, wo, wɨ/ never occur.
- **Do gradient patterns also exist?** Probably.
  ✓One potential candidate: Vowel-vowel sequences do occur but somewhat rarely. e.g., /ai/
- It is usually assumed that such phonotactic restrictions, categorical or gradient, and their strength differ depending on the **lexical strata**.
- Native and Sino-Korean words have different phonotactic patterns.
  e.g., Restricted occurrence of tense consonants in Sino-Korean. (Kwon 1997 etc.)

➢**I will explore phonotactics of native and Sino-Korean words, using UCLA phonotactic learner of Maximum Entropy model (Hayes and Wilson 2008).** cf. Cho (2012)

## 2. A maxent model of phonotactic learning

➢ **Conception: Quantitative pattern matching grammar**

- A maxent grammar assigns probabilities on phonological forms.
- The probabilities correspond to their phonotactic well-formedness.
- The model effectively detects rare but existing patterns.

➢**Characteristics**

- Only markedness constraints are learned.
- Inductive model: Constraints are learned without prior constraints.

➢ **Weighting on constraints by maximum entropy principle**

- To maximize the probability of the observed forms, the weights of constraints in a set Ω are assigned.
- Constraints with higher weights strongly restrict violated forms.

➢ **Searching constraints with heuristics**

- Accuracy: Observed/Expected ratio of constraints
- Generality: Shorter and general feature matrices are favoured.
- Under the thresholds of O/E, general constraints are selected.

## 3. Simulation

➢ **UCLA phonotactic learner** (Hayes and Wilson 2008) (http://www.linguistics.ucla.edu/people/hayes/Phonotactics)

➢ Training data: Common nouns including complex words

- Native-Korean: 6,121 words (from Cho 2002, Kang & Kim 2009)
- Sino-Korean: 22,859 words (from Kang & Kim 2009)
- Pronunciation forms based on Standard Korean dictionary (http://stdweb2.korean.go.kr/search/List_dic.jsp)

➢ All segments are not underspecified, except [+/-anterior].

## 4. Results: Constraints learned

### Categorical phonotactics

➢ Common, or similar, between native and Sino-Korean

| | | | |
|---|---|---|---|
| **Constraint 1:** | $*\begin{bmatrix}-\text{high}\\-\text{back}\\-\text{round}\end{bmatrix}\#$ | **meaning: */ɨ/#** | weight: 5.8 (Sino), 4.27 (native) |
| **Constraint 2-1:** | $*\begin{bmatrix}-\text{high}\\-\text{back}\end{bmatrix}\begin{bmatrix}-\text{sonorant}\\-\text{dorsal}\end{bmatrix}\#$ | **meaning: */ɛp, ɛs/** | Sino: weight 5.31 |
| **Constraint 2-2:** | $*\begin{bmatrix}-\text{high}\\-\text{back}\end{bmatrix}[-\text{sonorant}]\#$ | **meaning: */ɛp, ɛs, ɛk/** | native: weight 4.37 |

✓ C1: Words like loanword '스케이트 /sɨkʰeitʰɨ/' aren't attested in both lexicons.
✓ C2: Similar constraints are accidentally true for both lexicons
· Words like loanword '앱 [ɛp]' aren't allowed.

| ➢ **Sino-Korean only** | | **meaning** | **weight** | **cf. attested non-Sino-Korean words** |
|---|---|---|---|---|
| **C3:** | $*[+\text{aspirate}]\#$ | **No word-final aspirate** | 5.8 | 꽃 / k'ocʰ/ |
| **C4:** | $*\begin{bmatrix}-\text{sonorant}\\-\text{labial}\\-\text{dorsal}\end{bmatrix}\#$ | **No word-final coronal** | 5.84 | 낫 /nas/ |
| **C5:** | $*[+\text{tense}]\#$ | **No word-final tense** | 5.69 | 밖 /pak'/ |
| **C6:** | $*\#\begin{bmatrix}+\text{aspirate}\\+\text{dorsal}\end{bmatrix}[+\text{syllable}]$ | **No word-initial /kʰ + vowel/** | 4.54 | 코 /kʰo/ |
| **C7:** | $*\begin{bmatrix}-\text{round}\\-\text{syllable}\end{bmatrix}\begin{bmatrix}+\text{low}\\-\text{back}\end{bmatrix}$ | **No diphthong /jɛ/** | 4.54 | 얘기 /jɛki/ |
| **C8:** | $*[+\text{round}]\begin{bmatrix}-\text{sonorant}\\-\text{dorsal}\end{bmatrix}\#$ | **No word-final /op, up/** | 4.47 | 손톱 /sontʰop/ |
| **C9:** | $*\#\begin{bmatrix}-\text{high}\\-\text{low}\\-\text{back}\end{bmatrix}$ | **No word-initial /e/** | 3.45 | 에누리 /enuli/ |

➢ Cf. Previous studies
✓ C1, C3-6 and C9 are reported in the previous studies. (Kwon 1997, Kang 1998, An 2009, Shin 2009)
✓ C2 is from both lexicons. cf. A gap for Sino-Korean (Shin 2009)
✓ C8 is newly learned. It corresponds in part to */op, om, up, um/ reported in Kang (1998).

### Gradient phonotactics (i.e. constraints with exceptions)

➢ **Common**

- **C10:** $*\begin{bmatrix}+\text{high}\\+\text{back}\end{bmatrix}\begin{bmatrix}+\text{round}\\+\text{syllable}\end{bmatrix}$ **meaning: No /ɨ, u/ followed by /o, u/** weight: 4.08 (Sino) 3.24 (native)
  ✓ C10 learned in Cho's (2012) simulation

| ➢ **Sino-Korean only** | | **meaning** | **weight** | **exceptions** |
|---|---|---|---|---|
| **C11:** | $*\#[+\text{tense}]$ | **No word-initial tense** | 5.82 | words with 쌍 /s'aŋ/ |
| **C12:** | $*[+\text{syllable}]\begin{bmatrix}-\text{high}\\-\text{back}\end{bmatrix}$ | **No vowel followed by /e, ɛ/** | 4.39 | 차액 /cʰaɛk/, 우애 /uɛ/ |
| **C13:** | $*\begin{bmatrix}-\text{low}\\+\text{back}\\-\text{round}\end{bmatrix}[-\text{high}]$ | **No /ɨ, ʌ/ followed by non-high V** | 4.16 | 어업 /ʌʌp/, 저온 /cʌon/ |

| ➢ **Native-Korean only** | | **meaning** | **weight** | **exceptions** |
|---|---|---|---|---|
| **C14:** | $*[+\text{tense}]\#$ | **No word-final tense** | 4.53 | 밖 /pak'/ |
| **C15:** | $*\begin{bmatrix}-\text{sonorant}\\-\text{continuant}\\-\text{aspirate}\\+\text{coronal}\end{bmatrix}\#$ | **No word-final /t, c/** | 3.53 | 빛 /pic/ |
| **C16:** | $*\#\begin{bmatrix}-\text{high}\\-\text{back}\end{bmatrix}$ | **No word-initial /e, ɛ/** | 3.38 | 애벌레 /ɛpʌlle/ |
| **C17:** | $*\#\begin{bmatrix}+\text{high}\\+\text{back}\\-\text{round}\end{bmatrix}$ | **No word-initial /ɨ/** | 3.10 | 으뜸 /ɨt'ɨm/ |
| **C18:** | $*\begin{bmatrix}-\text{cont}\\+\text{asp}\\-\text{cor}\end{bmatrix}\begin{bmatrix}-\text{high}\\-\text{low}\\-\text{round}\end{bmatrix}$ | **No /kʰ, pʰ/ followed by /e, ʌ/** | 2.87 | 올케 /olkʰe/ |
| **C19:** | $*\begin{bmatrix}-\text{low}\\+\text{syl}\end{bmatrix}\begin{bmatrix}-\text{round}\\+\text{syl}\end{bmatrix}$ | **No high or medial V followed non-round V** | 2.79 | 헤엄 /heʌm/ |
| **C20:** | $*[+\text{tense}]\begin{bmatrix}-\text{low}\\+\text{back}\end{bmatrix}\#$ | **No tense preceding a word-final /u, o, ʌ/** | 2.66 | 대꾸 /tɛk'u/ |

➢ Hiatus avoidance constraints are active in both native and Sino-Korean lexicons.
✓ Relevant constraints: C10, C12-13, and C19
✓ Previous studies (e.g. Ha 2000): hiatus avoidance is active only in native Korean lexicon.
✓ But, 3 out of 4 constraints learned in the present simulation hold true for Sino-Korean lexicon.

## 5. Summary

- All categorical phonotactic patterns that have been reported in the previous studies were captured.
- Constraints for gaps and gradient patterns are newly learned.
- No categorical constraint was learned only from native-Korean lexicon.
- **The prediction of grammar will be examined by well-formedness test on nonce words.**

**Selected reference** An, S.-J. 2009. Characteristics of Sino-Korean word syllables. *Morphology* 11.1. 43-59. ◆ Berent, I, D. Steriade, T. Lennertz, and V. Vaknin. 2007. What we know what we have never heard: Evidence from perceptual illusions. *Cognition* 104:591-630. ◆ Cho, H-S. 2012. Statistical learning of Korean phonotactics. *Studies in Phonetics, Phonology and Morphology* 18.2: 339-370. ◆ Cho, N-H. 2002. *Hyentay Kwuke Sayongpinto Cosa: Hankwuke Haksupyong Ehwisencengul Wihan Kichocosa,* The National Institute of the Korean Language. ◆ Ha, S-K. 2000. Vowel hiatus resolution in Korean: An Optimality theory account. M.A. Thesis. Seoul National University. ◆ Hayes, B. and W. Colin. 2008. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry* 39.3: 379-440. ◆ Kang, B. and H. Kim (2009), *Hankwuke Sayong Pinto : 1500man Ecel Seyconghyengthayuymipwunsekmalmwungchi Kipan*, Hankwukmwunhwasa. ◆ Kang, Y-S. 1998. The Organization of lexicon in Korean. *Studies in Phonetics, Phonology and Morphology* 4: 55-67. ◆ Kwon. I-H. 1997. Hyentaykwuke hancaeuy umwunloncek kochal (The phonological study of contemporary Sino-Korean). *Kwukehak* 29: 243-260. ◆ Mikheev, Andrei. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics* 23:405-423. ◆ Shin, J-Y. 2009. Frequency related information and syllable structure constraints on Sino-Korean. *Malsoliwa Umsengkwahak* 1.2: 129-140.

**Acknowledgement** I appreciate Prof. Jun Jongho for invaluable advice on my research. Also, I give thanks to Jang Hayeon for providing training data and comments.

**e-mail: arimnet@naver.com**