Grammaticality and Lexical Statistics in Chinese Unnatural Phonotactics

Donald Shuxiao Gong

Grammaticality and lexical statistics in Chinese unnatural phonotactics

University of Kansas

gong@ku.edu

Speakers possess phonotactic knowledge about the acceptability of non-words, yet the source of this knowledge is unclear. One possibility is that a non-word is judged to be unacceptable because it violates the phonotactic grammar of this language. For instance, syllables in Standard Chinese take the form of CGVX (G=glide, X=vowel length, glide or nasal). To account for the syllable phonotactics of Chinese, four OCP-based phonotactic constraints can be proposed as part of such a phonotactic grammar, under the assumption that a natural phonotactic constraint is either 1) phonetically grounded, or 2) typologically well-attested (Hayes and White, 2013):

(1) Phonotactic Constraints in Chinese	Example
*HH: The feature [+high] cannot occur in sequence.	*[l**ui**] *[t**yu**]
*[Cor]_[Cor]: [Cor] cannot occur in both G and X.	*[**j**a**i**] *[p**j**e**i**]
*[Lab]_[Lab]: [Lab] cannot occur in both G and X.	*[**w**o**u**] *[n**w**a**u**]
C and G must have different articulators	*[**tʂj**an] *[**pw**aŋ]

Another possible account for the acceptability judgments is based on how similar the non-word is to all real words in the lexicon. Multiple models have been proposed to capture this analogical effect, and we focus on two of them: the Neighbourhood Density model (Bailey and Hahn, 2001) and Hayes & Wilson's Phonotactic Learner (Hayes and Wilson, 2008). Neighbourhood Density counts the number of words generated by substituting, deleting, or adding a single phoneme together with their summed frequency. For example, the form *lat* has abundant lexical neighbours in English (e.g. *cat*, *lap*), while *zev* has a sparse neighbourhood density. Phonotactic Learner produces a set of feature-based constraints given a feature matrix and a lexicon for training. The learner attempts to identify the constraint set and a set of constraint weights that maximise the probability of the input forms. We could then apply this learned grammar to evaluate the grammaticality of non-words by assigning penalty scores.

Linear logistic regression analyses were applied to the data of a phonological acceptability judgement mega study run on 110 Mandarin native speakers (Myers and Tsay, 2015). We used Neighbourhood Density, penalty scores generated by the Phonotactic Learner, and whether the phonotactic constraints in (1) are violated as independent variables to predict speaker's reaction time on the lexical decision task, with Neighbourhood Density and Phonotactic Learner representing lexical statistics, and constraints in (1) representing grammaticality. Results show that each parameter plays an independent role, suggesting that even though lexical statistics and grammaticality overlap substantially, each still independently contributes to speaker's reactions (Table 1). The results suggest that the extreme lexicalist view, which attributes all phonotactic patterns to frequency statistics (Hay, Pierrehumbert and Beckman, 2003) is too strong.

Non-words that violate the constraints in (1) are labelled as systematic gaps, while other missing syllables are labelled accidental gaps. However, some of the accidental gaps are not so 'accidental' as expected. We noticed a specific phonotactic constraint that bans the cooccurrence of a labial fricative with a following coronal glide (*[fj]), and incorporated it into the statistic model. Despite the constraint's phonetic unnaturalness, the reaction time results suggest that speakers reject *[fj] gaps more quickly than other accidental gaps, as if they were systematic gaps (Figure 1). Therefore, the relevance of this constraint in Chinese indicates that, unlike what has been proposed by Becker et al. (2011), unnatural phonotactics can be learned by speakers and be part of the phonotactic knowledge. The possibility that *[fj] is a natural

constraint or that it is a result of the phonemic analysis of Standard Chinese adopted here, however, will be discussed

	β	SE(*β*)	*z*	*p*
(Intercept)	-0.7212	0.0305	-23.730	
penalty	-0.0052	0.0020	-2.570	.0102*
neighbourhood density	0.0101	0.0028	3.578	.0003*
being a systematic gap	-0.0448	0.0212	-2.108	.0350*
penalty : *neighbourhood density*	-0.0007	0.0002	-3.093	.0020*

Table 1 Results of linear logistic regression on response

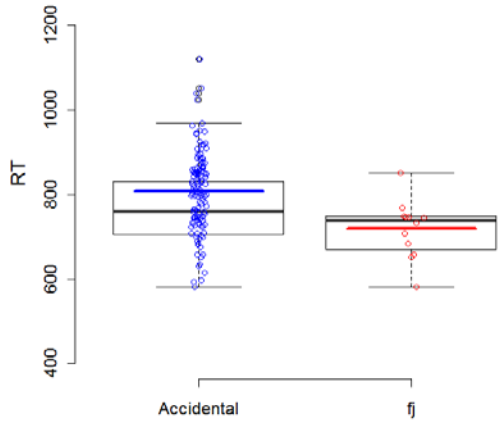

Figure 1 Reaction time distribution of accidental vs. *[fj] violating gaps

References

Bailey, T. M. and Hahn, U. (2001) 'Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods?', *Journal of Memory & Language*, 44(4), p. 568.

Becker, M., Ketrez, N. and Nevins, A. (2011) 'the Surfeit of the Stimulus: Analytic Biases Filter Lexical Statistics in Turkish Laryngeal Alternations', *Language*, 87(1), pp. 84–125.

Hay, J., Pierrehumbert, J. and Beckman, M. E. (2003) 'Speech perception, well-formedness and the statistics of the lexicon', in Local, J., Ogden, R., and Temple, R. (eds). Cambridge: Cambridge University Press, pp. 58–87.

Hayes, B. and White, J. (2013) 'Phonological naturalness and phonotactic learning', *Linguistic Inquiry*, 44(1), pp. 45–75.

Hayes, B. and Wilson, C. (2008) 'A maximum entropy model of phonotactics and phonotactic learning', *Linguistic Inquiry*, 39(3), pp. 379–440.

Myers, J. and Tsay, J. (2015) 'Mandarin Wordlikeness Project [Raw data]'. Available at: http://lngproc.ccu.edu.tw/MWP/index.html.