LEXICAL AND PHONOTACTIC EFFECTS ON WORDLIKENESS JUDGMENTS IN CANTONESE

James P. Kirby and Alan C. L. Yu

Phonology Lab, Department of Linguistics, University of Chicago
jkirby@uchicago.edu, aclyu@uchicago.edu

ABSTRACT

This paper reports the results of a wordlikeness task designed to investigate Cantonese speakers' phonotactic knowledge of systematic and accidental gaps. Regression analyses found that wordlikeness judgments correlate with token frequency-weighted neighborhood density and conditional (bigram) probability. This is suggested to be an effect of the relative phonological densities of the Cantonese and English lexica.

Keywords: phonotactics, wordlikeness, Cantonese, neighborhood density, conditional probability

1. INTRODUCTION

Early work in generative phonology focused on a distinction between possible versus impossible words [10]. Possible words became known as *accidental gaps*, while impossible words were presumed to be the result of *systematic gaps*. Classical generative phonology predicts speakers should categorically judge accidental gaps to be well-formed and systematic gaps to be ill-formed. However, research into well-formedness suggests speaker judgments of both words and nonwords to be gradient in nature, and influenced by lexical as well as phonotactic factors [1] [5] [6] [15].

With respect to gap well-formedness, hese observations suggest two hypotheses. First, if accidental gaps represent possible words and systematic gaps impossible words, speakers should rate accidental gaps more highly than systematic gaps in a wordlikeness judgment task. Second, if well-formedness is gradient, speaker judgments should differ within the group of systematic gaps; if the classical model is correct, no significant differences in judgments should be observed between systematic gaps. To this end, an experiment was performed probing Cantonese speakers' wordlikeness judgments on a set containing both possible and impossible words.

Since phonotactics often ban particular segments or sequences of segments, it was hypothesized that conditional phonotactic probability should correlate with gradient acceptability ('goodness' or 'wordlikness') of nonwords, as has been demonstrated previously [5] [6]. Lexical support has also been shown to correlate positively with nonword judgments [1] [15], leading to the prediction that neighborhood density would also correlate positively with nonword acceptability [9].

1.1. Cantonese phonology

Cantonese is a Yue language spoken in Hong Kong, Guangdong province of China, and by a large diaspora throughout the world. The language contains 19 consonants /p p^h t t^h ts ts^h k k^h k^w k^{wh} m n ŋ f s h l j w/, of which only nasals, glides, and /p t k/ are permitted in coda position. Cantonese rimes contain one of 8 monophthongs /aː a ɛː iː ɔː øː uː yː/ or 11 diphthongs /ai ɐi au ɐu ei ɛu ɵy ɔi ui iu ou/, and bear one of 6 tones /55 25 33 21 23 22/, with three 'checked' allotones /5 3 2/ occurring in syllables ending in unreleased /p t k/.

1.2. Cantonese phonotactics

This study considered three Cantonese phonotactic constraints: labial, coronal-vowel, and onset-tone restrictions, as well as so-called accidental gaps.

1.2.1. Labial gaps

Cantonese syllables are barred from containing both labial onsets and codas (**pap, *puːp*). Labial codas do not occur with rounded vowels (**-yːm, *-ɔːm*), nor do labial onsets occur with front rounded vowels (**møː-, *myː-*). However, the labial onset-coda restriction admits some exceptions in the form of loanwords and onomatopoeia [2].

1.2.2. Coronal-vowel gaps

Syllables containing both coronal onsets and codas together with the nuclei /ɔː uː/ are disallowed (e.g. **tɔːm, *tuːt*), as are syllables containing both coronal onsets and the nucleus /uː/ (**tuːp, *tuː*, etc.).

1.2.3. Onset-tone gaps

Unaspirated initials /p t ts k k^w/ do not occur in syllables with tones 23 or 21 (**pa23, *ta21*), and aspirated initials /p^h t^h ts^h k^h k^{wh}/ do not occur with the 22 tone (**p^ha22, *t^huː22*).

1.3. Previous analyses

Cantonese gaps, especially the labial gaps, have received some previous attention in the literature [2] [17]. Broadly speaking, these analyses attribute the gaps to Obligatory Contour Principle (OCP) violations. Work by Frisch and colleagues [7] [8] suggests that speakers of Arabic are sensitive to type and degree of OCP violation, and that this sensitivity manifests itself in wordlikeness judgments. When presented with novel verb forms which violate the relevant OCP-Place constraint, Frisch & Zawaydeh [8] found that judgments were gradiently influenced by consonant pair similarity. The authors also show that Arabic speakers rate systematic gaps much less wordlike than structurally equivalent accidental gaps. Similarly, Myers [13], Myers & Tsay [14], and Wang [16] report that speakers of Mandarin Chinese rate accidental gaps above systematic gaps in wordlikeness judgment tasks.

The extent to which Cantonese speakers behave like Mandarin and Arabic speakers in this regard was explored via the wordlikeness judgment experiment described in Section 2.

2. EXPERIMENT

2.1. Stimuli

As the goal was to investigate speaker judgments of different gap types, experimental stimuli were optimized for phonological simplicity. The stimuli consisted of a set of CV(C) words and nonwords, derived from all possible combination of the eight onset phonemes /f, p, p^h, m, s, t, t^h, n/, three vowel phonemes /aː, iː, uː/, an optional /m/ or /n/ coda, and six tones /55 25 33 21 23 22/. Velar onsets were omitted because initial /ŋ/ is often dropped by speakers. Final /p t k/ were also ignored because many speakers show final place neutralization that turn all /k/ into [t]. This method produced 432 syllables, of which 162 are attested and 270 are nonwords. Of the nonwords, 61 fill labial co-occurrence gaps, 36 fill onset-tone gaps, and 42 fill coronal gaps. 27 syllables filled two types of gaps simultaneously, and 1 syllable filled all three. The remaining 103 syllables were judged to be accidental gaps.

A male native speaker of Cantonese (the second author) was recorded producing in isolation the 33-tone version of each test syllable ($n = 72$). The six contrastive tones were resynthesized using Praat in order to avoid intra-syllable type variation.

2.2. Participants

Ten paid subjects participated in the experiment. All subjects were native speakers of Cantonese with varying levels of fluency in Mandarin and English. No subjects reported speech or hearing deficits.

2.3. Procedure

Subjects were presented with a randomized series of items from the corpus. For each stimulus, subjects were given two tasks. The first was to determine if the item is an existing word of Cantonese. The second was to rate the item for its wordlikeness on a 7-point scale, with 1 indicating "very poor - highly unlikely to be a real word of Cantonese" and 7 indicating "very good - a highly prototypical Cantonese word". There was no time pressure; subjects were allowed to repeat a given stimulus as many times as they wished, and were prompted to take a break after every 40 stimuli. Display and data collection were performed using the Praat software package.

2.4. Results

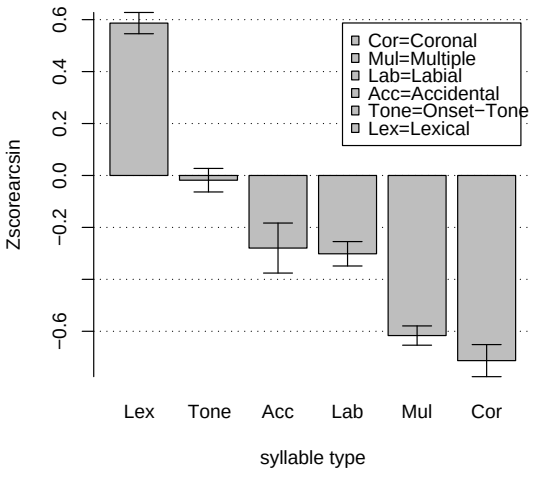
Individual wordlikeness ratings were scaled to the interval 0 to 1 and then transformed using the arcsine transformation (1):

$$x' = \frac{2 \arcsin\sqrt{x}}{\pi} \quad (1)$$

Transformed values for each subject were then normalized using a z-score transform and averaged over subjects.

As illustrated by Fig. 1, on average, all attested (lexical) syllables are rated above the baseline, while unattested syllables are rated at or below the baseline. However, wordlikeness ratings did not always favor accidental gaps over systematic gaps.

Figure 1: Mean arcsine-transformed goodness ratings by syllable type. Error bars show standard error for the mean.



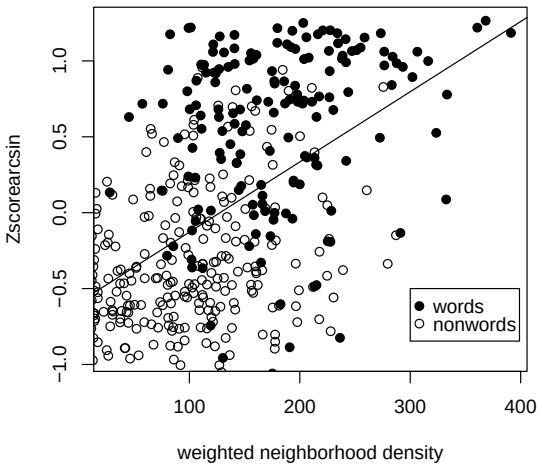
Post-hoc analyses using a Wilcoxon rank sum test determined the difference between wordlikeness ratings for words and nonwords to be significant ($U = 39217.5, p < 0.01$). Using Bonferroni adjusted α levels of 0.016 per test (0.05/3), mean wordlikeness judgments for accidental gaps were found to differ significantly from those of onset-tone (U = 1515, $p < 0.01$) and coronal (U = 3959.5, $p < 0.01$) gaps, but not those of labial gaps (U = 3319.5, p = 0.083). Using the same adjusted α level, differences in wordlikeness ratings were significant between labial and coronal (U = 480.5, $p < 0.01$), and coronal and onset-tone (U = 238, $p < 0.01$) gaps; although the difference between labial and onset-tone gaps did not reach the adjusted level of significance (U = 1405, p = 0.022), this may simply be a consequence of the small subject pool. Syllables filling more than one gap are not analyzed here.

Stimuli were also analyzed for the influence of phonotactic probability (PP) and neighborhood density (ND). Phonotactic probability (PP) for a stimulus was computed as the product of the item's component conditional bigram probabilities [5] [6]. Conditional probabilities were calculated based on both (lexical) type frequency and token frequency using frequency counts from the Hong Kong Cantonese Adult Language Corpus (HKCAC: [12]). Tonal probabilities were computed conditioned off of different syllabic constituents (onset, nucleus, coda), as well as ignoring tone entirely. The best results were achieved when conditioning tones off of nuclei and using bigram token frequencies; those results are reported here.

Neighborhood density (ND) was operationalized using a Levenshtein edit distance model [9]: ND for a word w was calculated by counting the number of lexical syllables in the Chinese Character Database [3] which could be formed by changing, adding, or deleting a single segment (or tone) of w. The contribution of a lexical syllable was weighted by its HKCAC token frequency, giving higher-frequency lexemes proportionally greater influence [14].

Multiple linear regression using ND and PP as factors captures a moderate amount of the overall wordlikeness rating variance ($R_a^2 = 0.3429$, F(2, 429) = 113.4, $p < 0.01$), with both factors emerging as significant regressors. Partial regression analyses reveal that a model using both regressors provides only a slight improvement over a model using ND alone ($R_a^2 = 0.3277$, F(1, 430) = 166, $p < 0.01$). Fig. 2 shows wordlikeness as a function of ND including the best-fit regression line. Perhaps unsurprisingly, attested syllables have greater mean neighborhood density ($\mu = 179.45, \sigma = 67.32$) than nonwords ($\mu = 96.57, \sigma = 63.07$).

Figure 2: Arcsine-transformed goodness ratings as a function of weighted neighborhood density.



Given that the difference in wordlikeness judgment values between words and nonwords reached significance, regressions were also performed on each group separately. For words, multiple regression using PP and ND as factors indicated a weak correlation ($R_a^2 = 0.05278, F(2, 159) = 4.43, p = 0.0134$). PP did not emerge as a significant regressor ($t = 0.67, p = 0.5$). For nonwords, a slightly stronger relationship is evident ($R_a^2 = 0.2144, F(2, 267) = 37.71, p < 0.01$), with both factors emerging as significant regressors. The model using both PP and ND provided a greater improvement over the ND-only model ($R_a^2 = 0.08805, F(1, 268) = 26.97, p < 0.01$) compared to the pooled results reported above.

Where present, the trends relating PP to wordlikeness ratings are weak, corroborating the regression analyses reported above. The exception to this generalization is the onset-tone gap, which shows a more significant trend ($R_a^2 = 0.2747$, $F(1, 34) = 14.26, p < 0.01$). For attested syllables, PP appears to correlate negatively with goodness ratings. ND is strongly correlated with wordlikeness for systematic gaps, particularly onset-tone gaps ($R_a^2 = 0.3724$, $F(1, 34) = 21.77, p < 0.01$); the trend is less pronounced for labial gaps ($R_a^2 = 0.2097$, $F(1, 59) = 16.92, p < 0.01$).

3. DISCUSSION

These results do not confirm the predictions of the classical generative phonology model, by which accidental gaps would be categorically well-formed and systematic gaps categorically ill-formed. Instead, this experiment found that some gaps consistently received higher goodness ratings than oth-

ers. Speakers appear to have clear judgments regarding the relative welll-formedness of zero-frequency syllables, and these judgments covary with type of gap filled: ND/PP and wordlikeness rating were more strongly correlated with judgments of onset-tone gaps than with those of coronal gaps, for example. In general, a correlation between ND and wellformedness is observed for both words and nonwords, with a less pronounced correlation between PP and wellformedness.

The relative weakness of PP as a correlate of well-formedness is unexpected in light of studies such as [6] which found that high-probability sequences had a positive effect on wordlikeness judgment for speakers of English. One possible explanation for the failure to observe this effect in Cantonese may have to do with the fact that English, which permits complex onsets and codas, allows for a far greater number of logically possible monosyllables ($n > 158,000$ [11]) than does Cantonese ($n = 5,130$ [19 initials $\times$ 45 rimes $\times$ 6 tones]). Crucially, English also makes use of a much smaller proportion of these possibilities than does Cantonese: the CMUDICT [4] lists around 10,000 non-homophonous English monosyllables (just over 6% of the logically possible combinations), whereas Cantonese has roughly 1,900 distinct syllables (around 36%). As a result, PP may emerge as a more important cue to wordlikeness in a language like English, where a smaller portion of the potential phonotactic space is occupied by lexical items.

This proportional discrepancy may also suggest an explanation for why neighborhood density correlates with wordlikeness in Cantonese to a greater extent than in English. Because it makes use of such a limited number of possible words, English nonwords which are routinely assigned high goodness ratings, such as *dresp*, often have no neighbors within a single segment edit distance. In Cantonese, where lexical items occupy nearly one-third the space of possible monosyllables, most nonwords have a neighbor within an edit distance of 1. These facts might offer an explanation for the relatively strong correlation observed in the Cantonese data between ND and wordlikeness ratings.

4. CONCLUSION

Our data show Cantonese speakers have gradient judgments regarding the ill-formedness of zero-frequency syllables. Regression analyses show a significant amount of the variability in judgments to be a function of token frequency-weighted neighborhood density and conditional phonotactic probability, affirming a role for frequency-based models in capturing at least some of the gradient wordlikeness judgment variance for zero-frequency sequences. When compared to similar studies for languages such as English, these results suggest language-specific roles for PP and ND with respect to their influence on the wordlikeness judgments of nonwords.

5. REFERENCES

[1] Bailey, T. M., Hahn, U. 2001. Determinants of wordlikeness: phonotactics or lexical neighborhoods? *J. Mem. Lang.* 44, 568–591.

[2] Cheng, L. L-S. 1991. Feature Geometry of Vowels and Co-occurrence Restrictions in Cantonese. *Proc. WCCFL 9*, 107–124.

[3] Chinese Character Database. http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/ visited 9-Feb-07.

[4] Carnegie Mellon University Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict visited 15-Mar-07.

[5] Coleman, J., Pierrehumbert, J. 1997. Stochastic phonological grammars and acceptability. *ACL SIGPHON* 3, 49–56.

[6] Frisch, S. A., Large, N. R., Pisoni, D. B. 2000. Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. *J. Mem. Lang.* 42, 481–496.

[7] Frisch, S. A., Pierrehumbert, J., Broe, M. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22, 179–228.

[8] Frisch, S. A., Zawaydeh, B. 2001. The psychological reality of OCP-Place in Arabic. *Language* 77, 91–106.

[9] Greenberg, J., Jenkins, J. 1964. Studies in the psychological correlates of the sound system of English. *Word* 20, 157–177.

[10] Halle, M. 1962. Phonology in Generative Grammar. *Word* 18, 54–72.

[11] Jespersen, O. Monosyllabism in English. *Proc. British Academy* 14, 341–368.

[12] Leung, M.-T., Law, S.-P. 2001. HKCAC: the Hong Kong Cantonese adult language corpus. *Intl. J. Corpus Ling.* 6, 305–326.

[13] Myers, J. 2002. An analogical approach to the Mandarin syllabary. *J. Chinese Phonology* 11, 163–190.

[14] Myers, J., Tsay, J. 2005. The processing of phonological acceptability judgments. *Proc. Symposium on 90-92 NSC Projects*, 26–45.

[15] Ohala, J. J., Ohala, M. 1986. Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In: Ohala, J., Jager, J. (eds), *Experimental Phonology*. Florida: Academic Press, 239–252.

[16] Wang, H. S. 1998. An experimental study on the phonotactic constraints of Mandarin Chinese. In: T'sou, B. (ed), *Studia Linguistica Serica*. Hong Kong: Lang. Info. Sciences Research Center, 259–268.

[17] Yip, M. 1989. Feature geometry and co-occurrence restrictions. *Phonology* 6, 349–374.