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
email: {jkirby,aclyu}@uchicago.edu

Phonotactic gaps

Accidental gaps: don't violate any phonotactic restrictions.
Systematic gaps: violate some phonotactic constraint(s).

Traditional grammatical approaches presume a *categorical* distinction between systematic and accidental gaps:

- all systematic gaps are equally ill-formed;
- all accidental gaps are equally well-formed.

This predicts *categorical* wellformedness judgments.

But: not all unattested words are judged identically!

- Acceptability of unattested words is *gradient*
- Acceptability reflected in *statistical properties of the lexicon* (n-gram probabilities, neighborhood density, etc.)

Previous studes focused on accidental gap acceptability, perhaps assuming systematic gaps are equally ill-formed [1] [2] [4] [6]

Research questions:

1. How do Cantonese speakers judge the wellformedness of systematic gaps?
2. Do the judgments correlate with lexical statistics?

Cantonese

(C)(G)V(V)(C) syllable structure
19 onsets: /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/
6 codas: /p t k m n ŋ/
8 monophthongs: /aː a ɛː iː ɔː øː uː yː/
11 diphthongs: /ai ɐi au ɐu ei ɛu ɵy ɔi ui iu ou/
6 tones: /55 25 33 21 23 22/

Typology of systematic gaps

- *Labial dissimilation gaps*
  - No labial onsets and labial codas (**pap, *puːp*)
  - No labial codas and rounded vowels (**-yːm, *-ɔːm*)
  - No labial onsets and front round vowels (**møː-, *myː-*)
- *Onset-tone gaps*
  - No aspirated onsets with 22 tone (**p^ha22, *t^huː22*)
  - No unaspirated onsets with 21/23 tones (**pa23, *ta21*)
- *Coronal gaps*
  - No coronal onsets and codas with /ɔː uː/ (**tɔːn, *tuːt*),
  - No coronal onsets with /u/ (**tuːp, *tuː*)

Experimental corpus

432 items conforming to a CV(C) template, derived from all possible combination of

- eight onset phonemes /f p p^h m s t t^h n/
- three vowel phonemes /aː iː uː/
- an optional /m n/ coda
- six tones /55 25 33 21 23 22/

Produces 162 attested syllables and 270 nonwords:

- 61 fill labial dissimilation gaps
- 36 fill onset-tone gaps
- 42 fill coronal gaps
- 27 syllables filled two types simultaneously, 1 all three

Remaining 103 nonwords classified as *accidental gaps*.

Procedure

Ten Cantonese native speakers were presented with a randomized series of items from the corpus & given two tasks per stimulus:

- *Lexical decision:* "Is this a word of Cantonese?" (y/n)
- *Wordlikeness rating:* "How good a word of Cantonese is this?" (1-7; 1 = worst, 7 = best)

Results

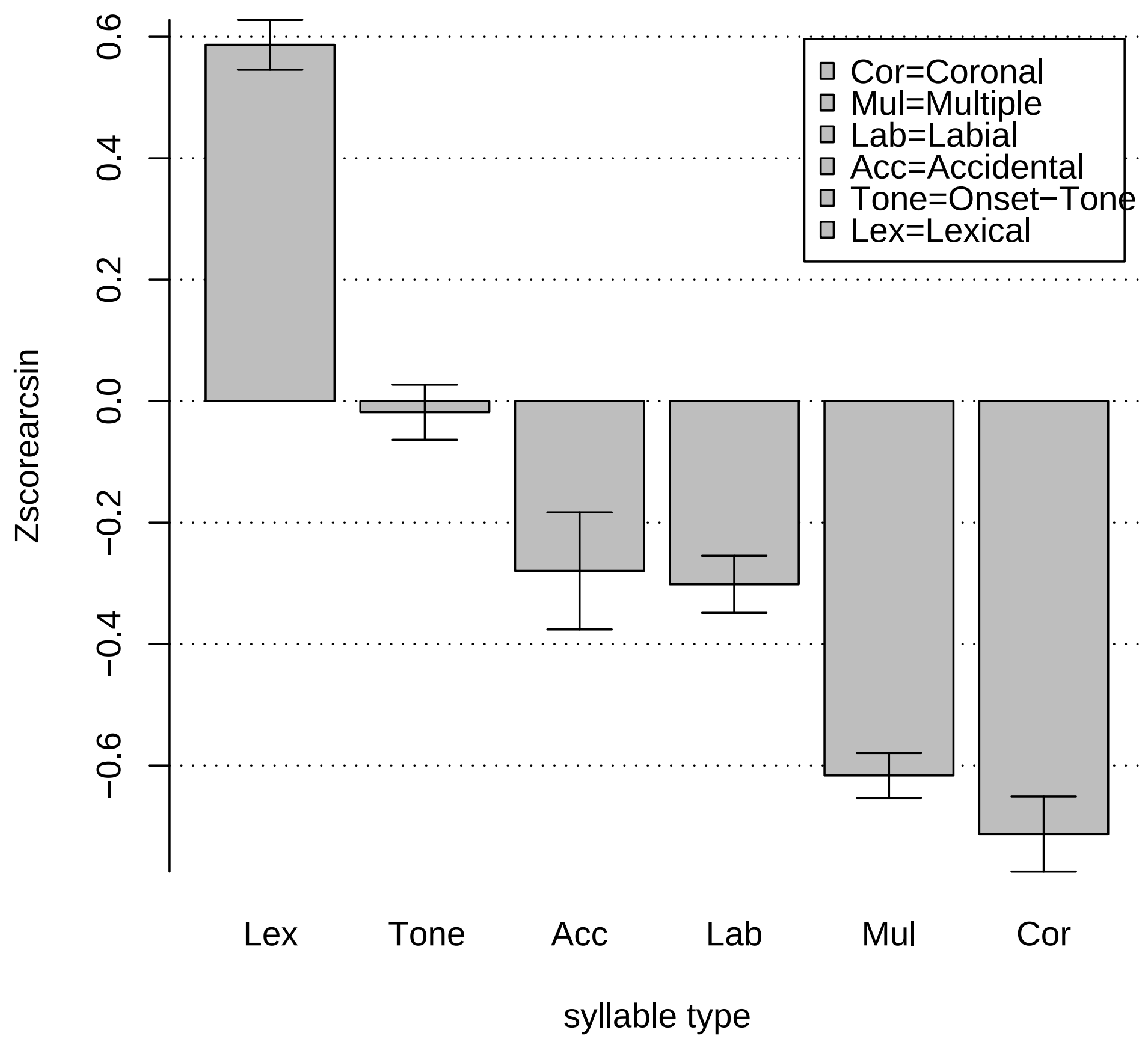
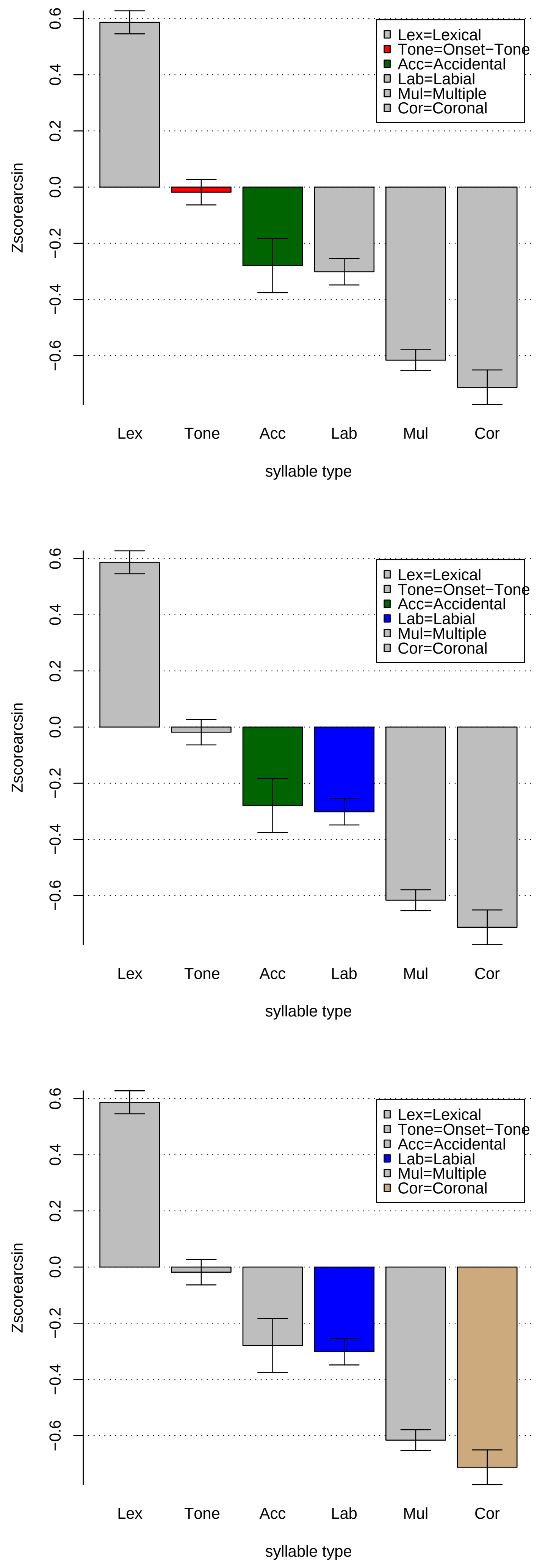


Figure 1: Mean arcsine-transformed goodness ratings by syllable type. Error bars show standard error for the mean.



Onset-tone
$\mu = 0.018$ $\sigma = 0.578$

Accidental
$\mu = 0.28$ $\sigma = 0.46$

Significant
$p = 0.0015$

Accidental
$\mu = 0.28$ $\sigma = 0.46$

Labial
$\mu = -0.302$ $\sigma = 0.367$

Not significant
$p = 0.0825$

Labial
$\mu = -0.302$ $\sigma = 0.367$

Coronal
$\mu = -0.713$ $\sigma = 0.24$

Significant
$p < 0.001$

Figure 2: Wilcoxon rank-sum results.

Lexical statistics

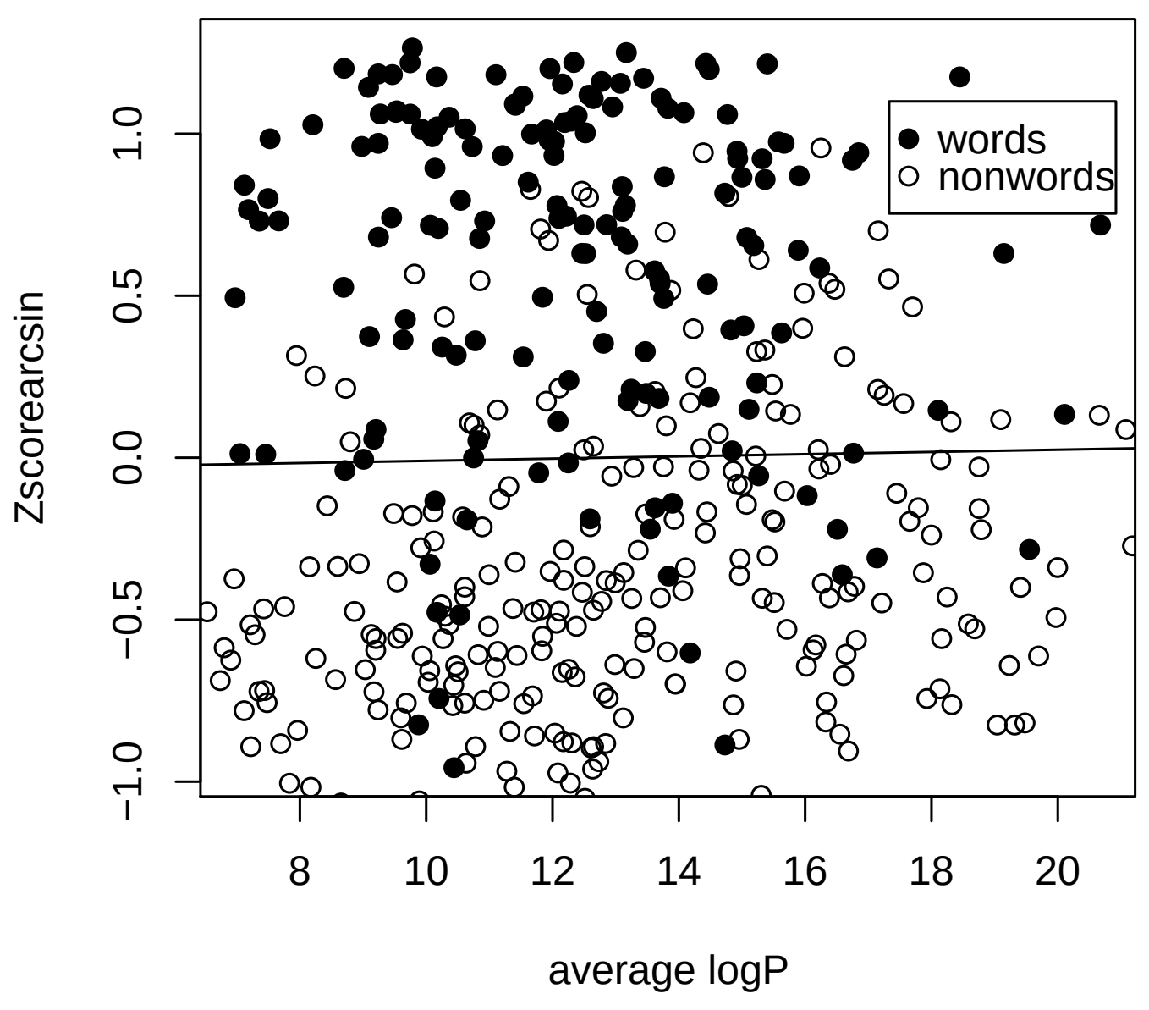
Phonotactic probability (PP) operationalized as average bigram log probability (1):

$$P(W) \approx \sum_{i=1}^{length(W)} -\log_2 p(w_i|w_{i-1}) \qquad (1)$$

Neighborhood density (ND) operationalized as Levenshtein edit distance between strings

ND(w) = number of syllables in the Chinese Character Database [3] which could be formed by changing, adding, or deleting a single segment (or tone) of w; weighted by token frequency in the Hong Kong Cantonese Adult Language Corpus (HKCAC: [5])

Results

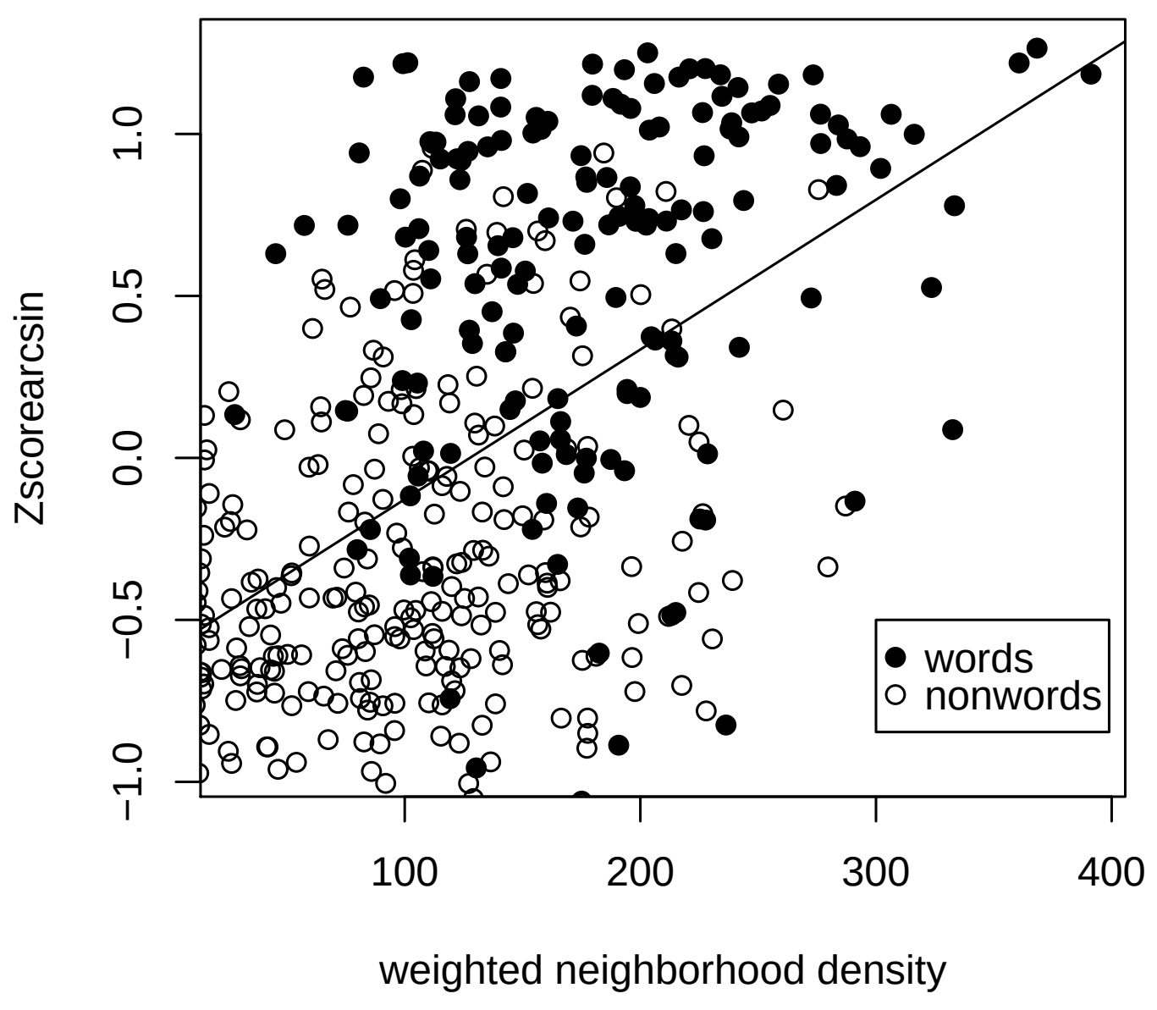


Not significant
$F(1,430) = 0.1387$ R^2_a: -0.002 $p = 0.71$

Words ●
$\mu = 12.34$ $\sigma = 2.79$

Nonwords ○
$\mu = 13.11$ $\sigma = 3.83$

Figure 3: Wordlikeness as a function of phonotactic probability by syllable type.



Significant
$F(1,430) = 166$ $R^2_a = 0.277$ $p < 0.001$

Words ●
$\mu = 179.45$ $\sigma = 67.32$

Nonwords ○
$\mu = 96.57$ $\sigma = 63.07$

Figure 4: Wordlikeness as a function of weighted lexical density by syllable type.

Subset	R^2_a	*df*	F	p	Factors
Words	0.052	2,159	4.43	$= 0.013$	ND
Nonwords	0.214	2,267	37.71	< 0.001	ND, PP
Both	0.343	2,429	113.4	< 0.001	ND, PP

Table 1: Multiple regression analyses.

Discussion

Our study found that speakers are sensitive to *degrees of ill-formedness among systematic gaps* and that their judgments *correlate with lexical statistics*, particularly ND.

Why is ND such a good predictor relative to PP? (cf. [4])

- English allows for a far greater number of logically possible monosyllables ($n > 158{,}000$) than does Cantonese ($n = 5{,}130$ [19 initials × 45 rimes × 6 tones])
- English also makes use of a much smaller proportion of the possibilities (10,000 monosyllables ≈ 6%) vs. Cantonese (1,900 monosyllables, ≈ 36%)
- For most Cantonese nonwords, ND(w) ≥ 1
- *The fact that most nonwords have lexical neighbors may underlie the emergence of lexical neighborhood density as a predictor of wordlikeness.*

Conclusions

- Gradient acceptability effects emerge even among nonwords which roundly violate phonotactic constraints.
- In Cantonese, acceptability seems to be correlated most strongly with lexical neighborhood density.
- Wordlikeness judgments are influenced by the phonotactic and lexical properties of a given language.

References

[1] Adam Albright and Bruce Hayes. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, 90:119–161, 2003.

[2] John Coleman and Janet Pierrehumbert. Stochastic phonological grammars and acceptability. In *Computation Phonology: ACL SIGPHON 3*, pages 49–56, Somerset, NJ, 1997. Assoc. Comp. Ling.

[3] Chinese Character Database. http://humanum.arts.cuhk.edu.hk/lexis/lexi-can/. Visited 09-Feb-07.

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

[5] Man-Tak Leung and Sam-Po Law. HKCAC: the Hong Kong Cantonese adult language corpus. *Intl. J. Corpus Ling.*, 6:305–326, 2001.

[6] John J. Ohala and M. Ohala. Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In J. J. Ohala and J. J. Jager, editors, *Experimental Phonology*, pages 239–252. Academic Press, Florida, 1986.

†**Address for correspondence: Department of Linguistics, University of Chicago, 1010 E 59th St., Chicago, IL 60637, USA.**