Non-Māori speaking New Zealanders show surprisingly sophisticated Māori phonotactic knowledge

Yoon Mi Oh, Clay Beckner, Jen Hay, Jeanette King
New Zealand Institute of Language, Brain and Behaviour, University of Canterbury

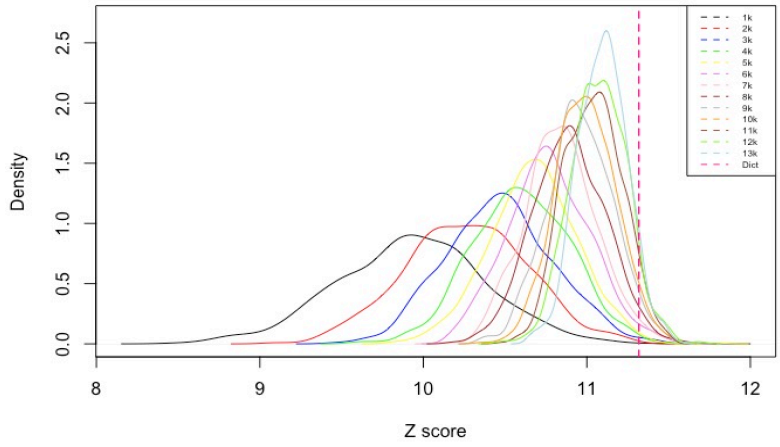
Language speakers can rate the gradient well-formedness of non-words in their language [1-2]. Such knowledge is assumed to have been acquired from statistical learning over speakers' lexicons [1-3]. Most New Zealanders (NZers) are exposed to Māori in their daily lives but do not speak Māori. They have a small lexicon comprising common loanwords and placenames. Our work has identified 121 words that most NZers can identify as Māori (although they can't define them all). We aim to understand the phonotactic knowledge that NZ-based non-Māori-speakers (NMS) & Māori-speakers (MS) have. Is it possible for non-speakers of a language to acquire statistical knowledge from only a small set of words?

Participants were asked to rate nonwords generated from a trigram model [4] for how good they would be as Māori words. We collected ratings for a total of 1760 words. Participants are 41 MS & 137 NMS. Phonotactic scores are calculated from: a Māori dictionary [5]; segmented Māori running speech data (RS) [6][7], unsegmented RS; & known words (the list of 121 words identified above, plus 55 placenames). Mixed-effects regression shows that both MS & NMS are influenced by phonotactics. The very best phonotactic predictor for both groups is the trigram model generated from the dictionary. There is no interaction between Māori-speaking status & the dictionary-derived phonotactic score. For both groups this is a much better predictor than phonotactics from known words, RS, or unsegmented RS, and both groups appear to be using these statistics equally as well. The phonotactic score derived from known words exhibits a significant difference between MS & NMS, such that the former appears to be less influenced by this than the latter. One possible reason that this is an inferior predictor to the dictionary is that the smaller size of the known-word set makes it a less robust training set. To further assess this question, we conduct Monte Carlo simulations using 1k random samples of 150 words from [5]. Our results again show that the known words are a better predictor than a random selection from the dictionary for NMS, but not MS, but the dictionary is the best predictor for both groups. Perhaps this is because a relatively small lexicon actually approximates the statistics of the whole dictionary.

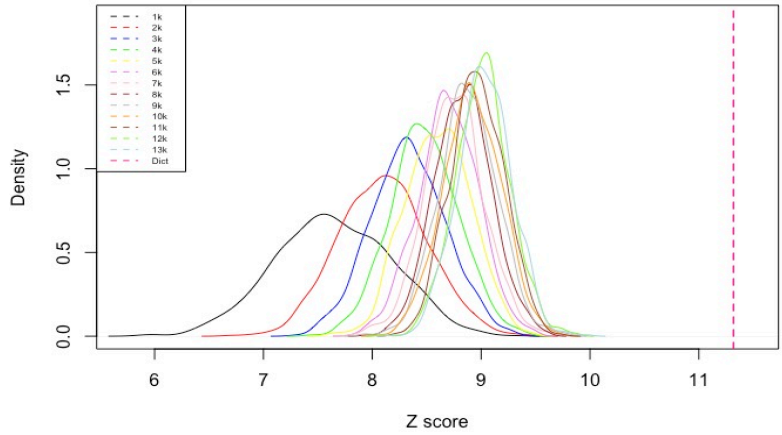
How big a lexicon would you need to have, to generate a similar trigram model to the dictionary-derived one? We tested Monte Carlo simulations with 1k random samples from [5] while varying the size of samples from 1k to 13k (c.f. Fig 1), using mixed-effects regressions to predict the ratings of NMS. As the vocabulary size increases, the z-scores of random samples gradually converge toward a full dictionary. This shows that NMS' actual phonotactic ratings of nonwords are successively better approximated by increasingly large samples of the Māori dictionary—with the very best predictions still provided by the full dictionary. This simulation appears to show that NMS' Māori phonotactic knowledge is best explained if we assume they have access to the full Māori lexicon. A similar set of simulations was done with frequency-rated random samples, with words selected into the individual samples in proportion to their lexical frequency. Phonotactics derived from these small lexicons perform still worse (c.f. Fig 2). Thus, we succeeded in our goal of demonstrating that speakers with a very small lexicon can have quite sophisticated phonotactic knowledge of a language. However we are left with the paradox that the knowledge appears to be too good to be true. To explain this result, we are exploring the possibility that the limited vocabulary of NMS provides enough initial phonotactics to allow for segmentation of ambient running speech, and this segmentation then leads to statistics derived from a much larger word-base than they appear to know. Regardless of the explanation, we have certainly found that non-speakers of a language can generate sophisticated phonotactic knowledge, and that phonotactics need not arise as a generalization over a large and established lexicon.

FIGURES

(1) Density plots from Monte Carlo simulation with different sizes of vocabularies (dictionary)



(2) Density plots from Monte Carlo simulation with different sizes of vocabularies (frequency-weighted dictionary)



Figures 1 & 2: Each colour shows a distribution of 1000 z-scores from mixed effects regression models, predicting NMS's ratings from a phonotactic score. The scores are generated by random samples of the dictionary (Fig 1) or by frequency-weighted random samples of the dictionary (Fig 2) to simulate vocabularies of different sizes.

[1] Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: effects of segment probability & length on the processing of nonwords. *Journal of Memory & Language*, *42*(4), 481–496. https://doi.org/10.1006/jmla.1999.2692.

[2] Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics & neighborhood activation in spoken word recognition. *Journal of Memory & Language*, *40*(3), 374–408.

[3] Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory & Language*, *62*(3), 311-331.

[4] Needle, J., Pierrehumbert, J. & Hay, J. (2014). Phonotactic probability & wordlikeness: a flexible pseudoword generator with triphones, Madison, WI, USA: *19th Mid-Continental Phonetics & Phonology Conference,* 12-14 Sep 2014.

[5] Moorfield, J. C., (2005). *Te aka: Māori-English, English-Māori dictionary & index,* Auckland, N.Z: Pearson Longman.

[6] King, J., Maclagan, M., Harlow, R., Keegan, P. & Watson, C. (2011). The MAONZE project: changing uses of an indigenous language database. *Corpus Linguistics & Linguistic Theory, 7*(1), 37-57.

[7] Boyce, M. (2006). *A corpus of modern spoken Māori.* PhD Thesis, VUW.