EVOLUTION OF PHONEME USAGE DRIVEN BY PHONETIC ROBUSTNESS FOR ERROR MINIMIZATION

INES PENA-NOVAS[1], MARCO ARCHETTI [*1]

[*]Corresponding Author: m.archetti@uea.ac.uk
[1] School of Biological Sciences, University of East Anglia, Norwich, U.K.

1. Introduction

Phonemes are, in principle, neutral with respect to meaning. Yet, it has long been known (at least since Dewey 1923) that certain phonemes are used more often than others by a 100-fold factor. What accounts for this phoneme usage bias?

Here we suggest that phoneme usage bias is driven in part by what we dub *phonetic robustness*, the capacity of a phoneme to reduce errors. While they are neutral with respect to meaning, phonemes have different probabilities of mutation (errors in articulation), and their mutants differ in their similarity with the original phoneme. Phonemes, therefore, differ in robustness — the capacity of reducing the probability of articulation errors and their effect on perception errors. These differences in robustness can lead to usage bias over time simply due to different probabilities of transmission. A similar effect has been studied in molecular evolution where synonymous codons (which are neutral at the protein level) are used with non-random frequencies because they differ in genetic robustness (Archetti 2004, 2006; Plotkin et al. 2004, 2006).

We propose a quantitative measure of phonetic robustness based on articulation and perception distances between phonemes; we show that phonetic robustness can lead to changes in phoneme usage over time in a deterministic theoretical model and in stochastic simulations; and we show that phonetic robustness can predict phoneme usage in English words.

2. Methods

Phonetic robustness. Robustness R_{p1} for phoneme $p1$ is the complementary value of the average of the perceptual distances $P_{p1,p2}$ from all phonemes p_2 weighted by the probability of mutation $(1\text{-}D_{\mathrm{p1,p2}})$ to p_2

DOI:10.12775/3991-1.091

$$R_{p1} = 1 - \sum_{p2} P_{p1,p2}\left(1 - D_{p1,p2}\right)$$

where $P_{p1,p2}$ is a normalised measure of perception distance (in this study, the distance between the first two principal components of phonological similarity – Mielke 2012) and $D_{p1,p2}$ is a normalised measure of articulation distance (in this study, the distance between the first two principal components of vocal tract distance – Mielke 2012).

Phoneme usage. Phoneme abundance was taken from the British National Corpus (Leech et al. 2001). Phonetic translation and phoneme frequencies were calculated using the Carnegie Mellon University Pronouncing Dictionary transcribed into IPA.

Theoretical analysis. The equilibrium frequencies of all phonemes were found by calculating the leading eigenvector of the matrix $(1\text{-}D_{p1,p2})P_{p1,p2}/\sigma_{p1}$, where σ_{p1} is a normalizing factor corresponding to the sum of the frequencies (before normalization) of the mutants of phoneme $p1$. We analysed a model with learning, in which errors can be corrected, and one without learning (σ_{p1}=1). We also analysed the same model in simulations for stochastic populations.

3. Results

Phonetic robustness is correlated with phoneme usage. We found a significant correlation between phoneme usage and robustness (R=-0.62, p<0.001 for all words). The correlation changes only slightly with the part of speech. A negative correlation means that less robust phonemes are used more often.

Phonetic robustness can lead to the observed phoneme usage bias. A model with learning, in which errors can be corrected, leads to an increase in frequency of the least robust phonemes. Phoneme frequencies change over time and their equilibrium values are correlated with robustness (R=-0.65, p<0.001) and with phoneme usage observed in the BNC (R=-0.57, p<0.005).

4. Discussion

Our results suggest that, if larger mutations can be detected and corrected more easily whereas mild mutations can persist undetected and uncorrected, robust phonemes will decrease in frequency over time (as their mutants are more likely to be transmitted), whereas anti-robust phonemes will persist (because their mutants are corrected, reverting to the original) and therefore increase in frequency over time. These results are in line with analogous observations in evolutionary genetics, where anti-robust codons in protein-coding genes increase in frequency over time because their mutants are detected and corrected with a higher probability than the mutants of robust codons. Our results suggest that phonetic robustness can explain why phonemes are used with unequal

frequencies is words, and therefore that phonetic robustness is a fundamental force driving the evolution of language.

References

Archetti, M. (2004) Selection on codon usage for error minimization at the protein level. *J. Mol. Evol.* 59, 400–415.

Archetti, M. (2006) Genetic robustness and selection at the protein level for synonymous codons. *J. Evol. Biol.* 19, 353–365.

Dewey, G. (1923), *Relative frequency of English speech sounds*. Harvard University Press, Cambridge, MA.

Leech, G., Rayson, P., Wilson, A. (2001) *Word Frequencies in Written and Spoken English based on the British National Corpus.* Longman, London, UK.

Mielke, J., (2012) A phonetically based metric of sound similarity. *Lingua* 122, 145–163.

Plotkin, J.B. Dushoff, J., Desai, M.M., Fraser, H.B., (2006) Codon usage and selection on proteins. *J. Mol. Evol.* 63, 553–635.

Plotkin, J.B. Dushoff, J., Fraser, H.B. (2004) Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. *Nature* 428, 943–945.