CONTRASTING PHONETIC COMPLEXITY ACROSS LANGUAGES: TWO APPROACHES

Caleb Everett [*1]

[*]caleb@miami.edu
[1]Anthropology, Psychology, University of Miami, USA

This paper examines phonetic complexity via two approaches that rely on transcribed word lists. Both approaches focus on results obtained for a 37-language sample, but contrast these results with findings from about 7000 language varieties. For the first approach, complexity is measured simply as the ratio of types of phones to tokens of phones, for each list representing a particular language variety. The second approach operationalizes complexity as unpredictability of sound usage, and simplicity as predictability. Predictability is based on the global mean frequency of occurrence of 41 sound types across all language varieties in the data. These global frequencies are then used to predict sound usage in the 37 languages focused upon, with less predictable languages deemed more "complex". Three languages in the sample are found to be complex according to both metrics explored here, while two languages are found to be simple according to both metrics. These findings are exploratory given the limitations of the word lists tested.

1. Introduction

Languages vary markedly in terms of the number of sounds they utilize. One could argue that languages with more phonemes represent complex phonological systems, though such a claim overlooks non-phonemic parameters including syllable structure and prosodic phenomena. Still, we can speak of specific kinds of complexity, e.g. complexity of phonemic inventories, without making presumptions regarding overall phonological, phonetic, or otherwise linguistic complexity. In this study I offer two approaches to looking at the complexity of languages' variant usage of sounds, both of which focus upon the phonetic units in basic transcriptions of 40-100 words (Swadesh-type lists). I apply both methods to the 37-language sample but, as critical background to this sample, I also apply the metrics to thousands of other languages.

2. Type:token ratio of phonetic segments

The first metric of complexity is simply the type:token ratio of transcribed phonetic units. The assumption underlying this metric is that languages with a greater density of sound types are more complex in terms of their sound-type inventories. I say "sound-types" as opposed to phonemes because this study relies on the ASJP database, a collection of roughly 7000 word lists that are phonetically transcribed. The phonetic transcriptions in the database are somewhat broad, as they use 41 basic sound types (Wichmann et al. 2016). Still, despite any limitations, there are advantages to using a database representing so many languages, as we can contrast our results for the 37-language sample with results from the bulk of the world's languages. (Over 4500 distinct ISO codes are represented in the data.)

To calculate the type:token ratio, I simply summed the number of unique sound types represented in a word list, and then divided that sum by the total number of sound tokens represented in the list. Secondary symbols for nasalization and other phenomena were ignored. Since this study focused on phonetic segments as opposed to phonemes, two-sound sequences such as prenasalized stops were treated as separate sounds. To contextualize the type:token ratios obtained for the 37-language sample, I gathered type:token ratios for about 7000 other varieties in the database. (I excluded varieties for artificially constructed languages.) I then obtained family-level averages of these ratios. The 264 family groupings were based on the WALS database (Dryer et al. 2013). Family means of type:token ratios ranged from 0.026 to 0.283. The overall mean across families was 0.121. The mean for the 37-language sample was about the same, at 0.119. (For a list of all family means, see the supplemental material.) The following ordering was observed, for the 37-language sample, from highest to lowest type:token ratio: 1. Norwegian (Nynorsk) 2. Catalan 3. Portuguese 4. Afrikaans 5. Danish 6. Arabic 7. Swedish 8. Polish 9. Czech 10. Slovak 11. Slovenian 12. Urdu 13. Turkish 14. Hebrew 15. Dutch 16. Galician 17. Croatian 18. Romanian 19. Italian 20. Norwegian (Bokmaal) 21. Bulgarian 22. Ukrainian 23. Vietnamese 24. Latvian 25. Mandarin 26. Greek 27. English 28. Hungarian 29. French 30. Persian 31. Hindi 32. Estonian 33. Russian 34. Serbian 35. Finnish 36. Spanish 37. Basque (See results file.)

To be clear, the suggestion being made here is not that languages with higher type:token ratios are necessarily more complex in terms of articulation. I am simply proffering one way of exploring phonetic segment complexity, one that

could be tested for associations with socioecological factors. This approach could also be applied to more robust intra-linguistic samples.

3. Predictability of sounds' usage rates

Another way to think of phonetic complexity is in terms of deviation from a typologically based expectation of languages' usage of individual sound types. According to such an approach, languages that use crosslinguistically uncommon sounds frequently, or common sounds very infrequently, would be more unpredictable and therefore more "complex" in a typological sense.

Given the lists of sounds in a particular word list, we can predict (roughly) how much each sound is used (Everett, under revision). For instance, we may predict that an alveolar nasal is used frequently, a voiceless alveolar stop a bit less so, a voiced alveolar stop even less, and so on. (Assuming these sounds are all present in the language in question.) The second metric for complexity adopted here relies on the fact that sounds' "usage rates" are somewhat predictable. Usage rates refer to the proportion of all the sound tokens in a given word list that are represented by a given sound. For instance, if there are four tokens of [t] in word list, out of 400 total sounds in the words in the list, then the usage rate of [t] is simply 0.01. Usage rates can be used to test the predictability of the occurrence of sounds across the world's languages. To do so, I adopted the following five steps: 1) Usage rates were obtained for all 41 sounds in each of the 6902 language varieties tested. 2) The average family-level usage rates were found for all sounds for each of 264 WALS language families. 3) These family-level averages were then averaged, resulting in phylogenetically controlled average usage rates for all sounds. 4) The sounds were then ranked according to their usage rates, at a global scale. (Sound rankings and mean usage rates are presented in the supplemental material.) 5) These global rankings were used to generate the predicted usage rates of sound types for individual languages, and these predicted usage rates were then contrasted with actual usage rates. Step 5 requires some elaboration: How are sound rankings, from most (#1) to least (#41) used in the world's languages, transformed into predicted usage rates? I transformed the rankings into predicted usage rates via the Borodovsky and Gusein-Zade formula. This formula was developed to predict the frequency of phonemes within a language from the frequency ranking of phonemes for that language (Tambovtsev and Martindale 2007). The formula allows us to predict a phoneme's intralinguistic frequency (f_r) from its intralinguistic rank *(r)*:

$$f_r = \frac{1}{n}\,(log(n+1) - log\ r)$$

For this study, I used the same formula but used the crosslinguistic rank, arrived at in step 4 above, as *r*. For *n*, I used the number of basic sounds in the database, or 41. (I should note that this formula only provides a marginal improvement to simply using the observed global average usage rates as the expected usage rates within each language, without using any transformation at all.)

With the predicted usage rates of the 41 sounds in hand, I then focused on the actual usage rates of the sounds in the 37-language sample. The association between predicted and actual usage rates, for all 37 varieties, is depicted in Figure 1. For each of the 37 languages, I ran a regression testing the association between predicted usage and actual usage. Higher R^2 values correspond to greater overall predictability. R^2 values ranged from 0.76 to 0.15, with a median of 0.51. Lower R^2 values are suggestive of greater usage-based deviance from a crosslinguistc norm, a kind of typologically-based complexity. The association is quite robust in most cases, but some varieties are clearly more predictable vis-à-vis their usage of sounds in these word lists. (See Figure 1 below. See results file for R^2 values of each of the 37 languages.)

4. Conclusion

I have outlined two potential approaches, of many, to measuring phonetic complexity. Each approach is based on a different interpretation of what is meant by complexity. One considers languages with predictable usage rates to be less complex (though admittedly this operationalization equates typologically anomalous usage with complexity, a strategy open to debate), the other considers languages that rely repeatedly on the same sounds, with relatively sparse usage of distinct sound types, to be less complex. The metrics resulting from these approaches are admittedly coarse but, I think, useful as exploratory measures. The complexity rankings of the 37 languages are somewhat similar for both metrics (Spearman's rho=0.44, p=.007). Finally, some remarks on individual languages: Interestingly, three closely related languages are the three most complex languages according to the predictability metric: Swedish, Danish, and Norwegian (Bokmaal), in that order. Swedish and Danish are also in the top 7 according to the type:token metric. Norwegian (Bokmaal) is not amongst the most complex according to the type:token metric, though Norwegian (Nynorsk) is. So there is some Scandinavian flavor to the more complex varieties according to both metrics, but also some cross-dialectal variability (which admittedly may simply be the artifact of the small sample

sizes). In contrast, both Finnish and Basque are ranked amongst the three "least complex", for both metrics. Of course it remains to be seen just how much these findings, based as they are on short word lists, are representative of larger patterns in these languages. What we can say is that, given the metrics and data utilized here, there is some observable though modest coherence at both ends of the range of complexity, for this sample of 37 languages.

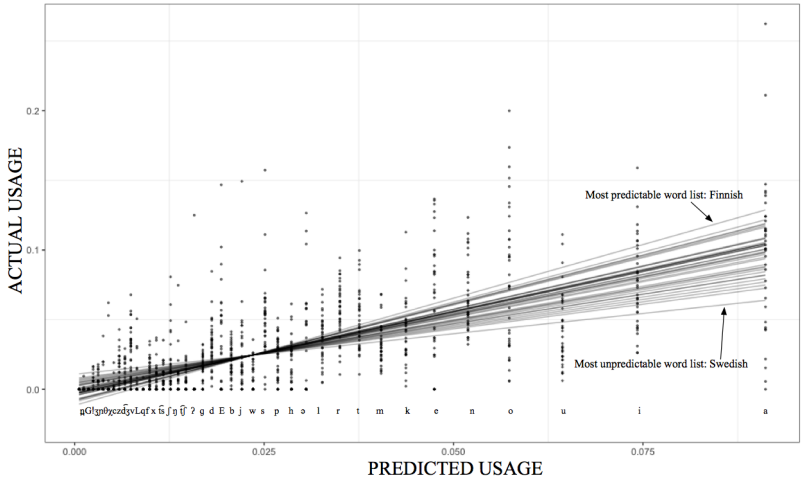

Figure 1. Relationship between predicted usage and actual usage, for each of the 37 languages in the sample. Each LMS line depicts the association between the typologically based predicted usage of 41 sound types and a language's actual usage of those sound types (judging from the transcribed word lists). Each column of 37 dots represents the usage rates of one of the 41 sounds, across each language in the 37-language sample.

References

Everett, C. (Under revision) The predictability of sound usage across languages.

Dryer, Matthew S. & Haspelmath, Martin (eds.) (2013) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Tambovtsev, Y. & Martindale, C. (2007) Phoneme frequencies follow a Yule distribution: The form of the phonemic distribution in world languages. *SKASE Journal of Theoretical Linguistics* 4.2

Wichmann, S., Holman, E., and Brown, C. (2016) The ASJP Database.