Sonority principles predicting cross-linguistic patterns in phonotactics also predict within-language probabilistic distributions of segment sequences

Peiman Pishyar-Dehkordi [1]
[1] *Linguistics Department, University of Canterbury (New Zealand)*

The worlds' languages contain a variety of cross-linguistic phonotactic patterns, in which certain sound sequences are more likely to occur in languages than others [1]. We investigate whether these phonotactic patterns are reflected as probabilistic distributions within individual languages in the context of sonority constraints on syllable formation.

Cross-linguistic patterns in syllable structure have been argued to be governed by sonority hierarchies [2,3]. Sonority hierarchies are loosely based on the comparative loudness of speech sounds. The Sonority Dispersion Principle (SDP) argues that: "… the simplest syllable is one with the maximal and most evenly-distributed rise in sonority at the beginning, and the minimal drop in sonority at the end. Syllables are increasingly complex to the extent that they depart from this preferred profile." [2]. Simpler syllables as defined by the SDP are more cross-linguistically common (attested in more languages) than complex syllables. For example, plosive + vowel syllables are attested in more languages than glide + vowel syllables based on SDP [2].

We also know that, in addition to categorical constraints on their phonology, languages also contain probabilistic phonotactic patterns within the sequences that they allow [4,5]. Amongst their legal phonological sequences, some sequences tend to be over-represented, and some sequences under-represented. Native speakers of languages can rate the well-formedness of non-words, closely tracking the probabilities of the sequences they contain [6,7]. These within-language gradient phonotactic patterns are thus an important part of linguistic knowledge.

If cross-linguistic patterns are also reflected as within-language probabilistic constraints, we can hypothesize that in a language that allows both simple and complex syllables based on the universal sonority dispersion principle, simpler syllables will be more frequent within that language. This hypothesis generates a number of untested predictions, for example, that in a language that allows for both plosive + vowel and glide + vowel syllables, plosive + vowel syllables will be more frequent than glide + vowel syllables.

To test this hypothesis, we have used databases of phonemically transcribed lemmas with syllable boundaries to obtain the type frequency of bigrams in 4 positions of #[CV], [VC]#, #[CC]V and V[CC]# within syllables in 8 languages: English, German, Dutch, te reo Maori, Italian, Portuguese, French and Greek. Linear regression models were performed to test any correlation between cross-linguistic sonority-based complexity of bigrams and within-language probabilistic distribution of bigrams. In order to make sure the model predictions are far from random, Monte Carlo simulations were performed by generating 1000 random iterations of bigram lists of each language and performing the similar statistical models on each of 1000 random iterations to see whether the model predictions for the original lists are significantly different from those for the random lists.

Results are consistent with our hypothesis in that a significant correlation exists between bigram complexity based on the sonority dispersion and bigram type frequency for #[CV], [VC]# and #[CC]V positions within syllables (less complex is correlated with more frequent) (Table 1 shows Monte Carlo results using shades of colour). These findings suggest that in languages explored in this study, there is a general alignment between sonority dispersion principle as a cross-linguistic universal and within-language probabilistic distribution of segment sequences within syllables.

Table 1. The numbers represent the probability of occurrence of the original bigram lists within 1000 random iterations of them. Shades of green are used when the probability is in support of our hypothesis. Shades of red are used when the probability is against our hypothesis.

	#[CV]	[VC]#	#[CC]V	V[CC]#
Dutch	0.999	0.999	0.000	0.000
French	0.001	1.000	0.000	-0.009
Italian	0.000	1.000	1.000	0.997
English	0.000	0.000	0.000	1.000
German	0.000	0.428	0.005	-0.001
Greek	1.000	0.000	0.000	0.823
Portuguese	0.000	0.000	0.000	-0.021

References

[1] Kenstowicz, M. J. (1994). *Phonology in generative grammar* (Vol. 7). Blackwell Cambridge, MA.

[2] Clements, G. N. (1990). The role of the sonority cycle in core syllabification. *Papers in laboratory phonology*, *1*, 283-333.

[3] Clements, G. N. (1992). "The Sonority Cycle and syllable organization." In Wolfgang U. Dressler, Hans C. Luschützky, Oskar E. Pfeiffer, and John R. Rennison, eds. Phonologica 1988: Proceedings of the 6th International Phonology Meeting. Cambridge: Cambridge University Press. pp. 63-76.

[4] Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural language & linguistic theory*, *22*(1), 179-228.

[5] Pierrehumbert, J. (2001). Stochastic phonology. *Glot international*, *5*(6), 195-207.

[6] Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods?. *Journal of Memory and Language*, *44*(4), 568-591.

[7] Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of memory and language*, *42*(4), 481-496.