

## Phonological Similarity and Identity

Effects of phonological similarity on language processing have received an enormous amount of attention in psycholinguistics, as they provide a window into processes underlying lexical production [1,4-7]. Despite the fact that similarity has also been extensively studied in work on typology, there has been little cross talk between the disciplines. We present a large-scale cross-linguistic study of the role of similarity in the attestation of theoretically possible word forms in the mental lexicon. The striking parallels we find suggest a deep causal link between effects observed during incremental production (speech errors, speech rate, fluency) and typological patterns. Our findings also stress the importance of distinguishing similarity from identity, which are often conflated in psycholinguistic work.

Using a typologically diverse set of 11 languages (Aymara, Chol, Dutch, English, French, German, Greek, Muna, Quechua, Turkish, Zulu), we compare observed word forms against theoretically possible ones given a language's phoneme inventory. For every possible bi-consonantal template (e.g.,  $C_1VC_2$ ), we coded *Identity* (whether  $C_1$  and  $C_2$  were identical) and *Similarity* (according to [2]). Note that under this coding, the identity effects described below express the effect of identity in addition to the maximal similarity of identical consonants.

**Study 1a** employed Poisson-regression to assess effects on the popularity of a template (how often a template actually occurred in a language) controlling for the independent frequency of consonants in position  $C_1$  and  $C_2$ . As expected, this control accounts for the bulk of the variance in popularity (Fig.1, gray bars; collinearity controlled). We found strong similarity avoidance, but an affinity for identity. Similarity had a strong negative effect in all languages (all p-values  $<.0001$ ; Fig.2, blue bars). Identity improved models for all languages except German (all other p-values  $<.005$ ). Where significant, identity always had a positive effect: templates with pairs of identical consonants are more frequent than expected from their similar (Fig.2, red bars). Importantly, however, identical consonants are nonetheless under-attested given their independent occurrence probabilities, in all languages except Zulu (Fig.2, green bars).

**Study 1b** confirmed that results hold separately for attestation (whether a template occurs at all in the language) and popularity (frequency of a template excluding unobserved templates), suggesting that both categorical and gradient typology reflect identity and similarity biases.

**Study 2** investigated whether the relative strength of similarity effects is uniform across phonological features. We find that certain features (in particular [+labial]) induce much stronger similarity avoidance than others (e.g. manner features). Our findings bear a striking resemblance to work on speech errors [7].

**Study 3** extended our work to tri-consonantal templates (e.g.,  $C_1VC_2VC_3$ ). Crucially, we show that some aspects of similarity and identity effects are position dependent (cf. [5] for similar findings from speech rate).

We find that languages across the world exhibit a strong bias against segmental similarity that we link to interference during language production. Our findings call for more typologically informed work on production (and vice versa, cf [3]). We discuss the consequences and necessary extensions to models of language production.

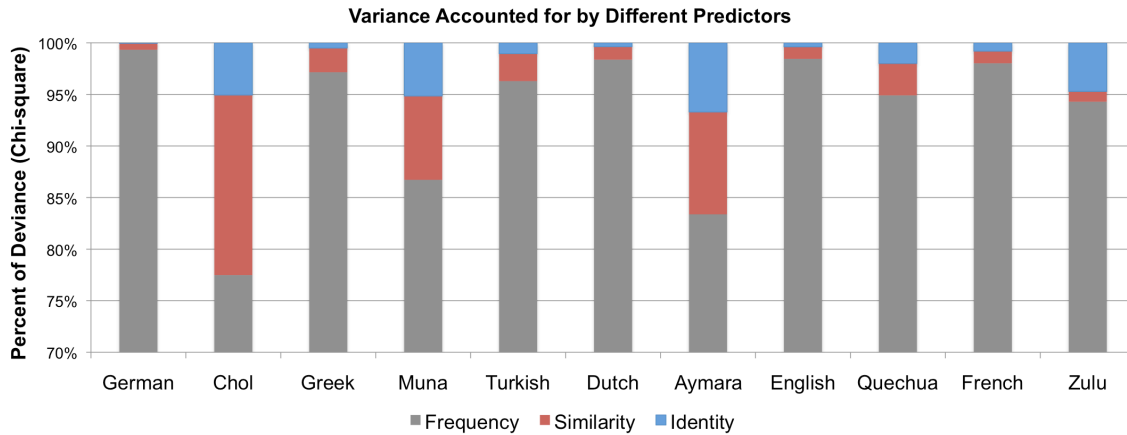


Figure 1 – Variance Accounted for by Identity, Similarity and Frequency Predictors

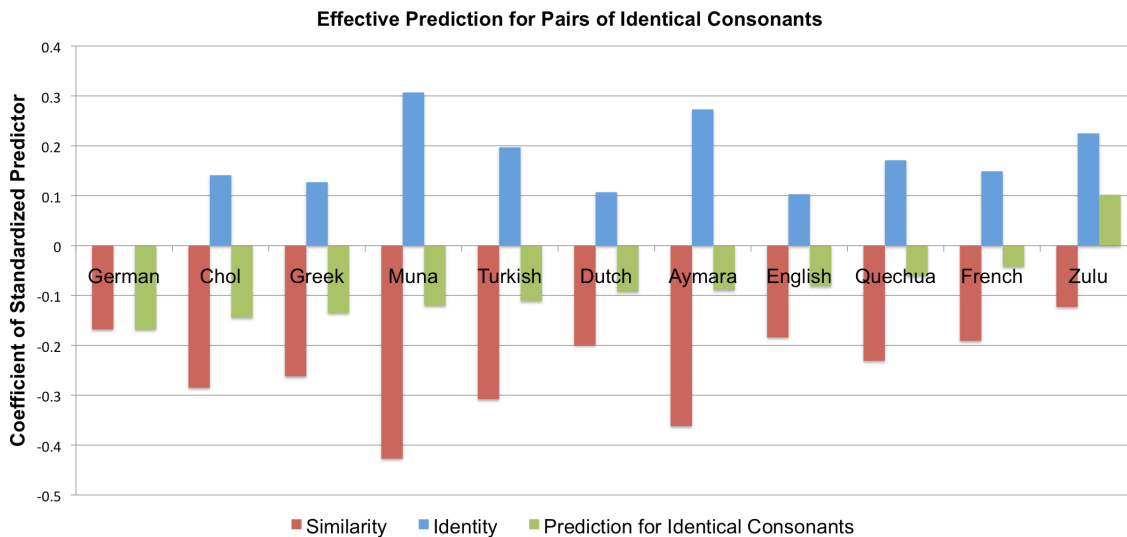


Figure 2 – Coefficients of Similarity Predictors, Identity Predictors and Model Prediction for Identical Pairs of Consonants

**References.** [1] Dell, G. (1986) A spreading activation theory of retrieval in sentence production, *Psychological Review*, 93, 283–321. [2] Frisch, S., Pierrehumbert, J. and Broe, M. (2004) Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22, 179-228. [3] Hawkins, J. A. (2007). Processing typology and why psychologists need to know about it. *New Ideas in Psychology*, 25, 87–107. [4] Janssen, N., & Caramazza, A. (2009). *Grammatical and phonological influences on word order*. *Psychological Science*, 20: 1262-1268. [5] Sevald C., Dell G. (1994) The sequential cuing effect in speech production. *Cognition*. 53:91–127. [6] Starreveld, P. (2000). On the interpretation of auditory context effects in word production. *Journal of Memory and Language*, 42, 497-525. [7] Stemberger, J. P. (1991). Radical underspecification in language production. *Phonology*, 8, 73 – 112.