# Initial syllable faithfulness as the best model of word-size effects in alternations

Lexically-specific phonological alternations are now well-known to have skewed distributions in the lexicon: languages rarely, if ever, have 50% of their words undergoing an alternation (Ernestus & Baayen 2003). In Turkish, word-final stops become voiced upon the addition of a vowel-initial suffix (*amatʃ* ~ *amaʤ-ɨ* 'goal') in some nouns, while other nouns (*atʃ* ~ *atʃ-ɨ* 'hunger') resist the alternation (Lees 1961 et seq.). Whether a given noun undergoes this alternation depends on *word size*: short nouns resist the voicing alternation more than longer nouns. The literature offers three different types of accounts for the effect of word-size on the distribution of alternations within the lexicon: (a) syllable-based, where initial syllables enjoy special faithfulness protection, (b) mora-based, where bi-moraic minimality protects CVC words from alternations, and (c) lexical-neighborhood-based, where a word's neighbors (Luce & Pisoni 1998) inhibit its propensity to undergo an alternation. In this paper, we offer a statistical study of Turkish and its mirror image, Russian, which shows that the syllable-based account offers the best description of the variability in the data. We use this result to motivate an initial-syllable faithfulness analysis of the word-size skewing of final voicing alternations.

In this paper we compare three accounts of the word-size effect in Turkish: (a) Becker et al. (2008) claimed that in monosyllabic roots, final stops are protected by initial-syllable faithfulness, but in polysyllabic roots they aren't, while (b) Inkelas & Orgun (1995), Inkelas et al. (1997), and Pycha et al. (2007) described the size effect as a difference between minimal CVC roots (where a final consonant is syllabified early to achieve bimoraic word minimality) and anything longer, grouping CVCC roots with poly-syllables. Finally, (c) Wedel (2002) argues that the real source of the size effect is neighborhood density; according to this explanation, shorter words have more neighbors, pressuring them to keep their shape constant in order to facilitate lexical access in the face of many phonologically-close lexical competitors.

We judged the three explanations by their R^2 scores, a statistical measure of variance. The chart in (1) shows the goodness of the three theories as predictors of voicing alternations in a lexicon of Turkish (TELL, Inkelas et al. 2000). Logistic regression models were fitted to the data in 1000 bootstrap samples for each of the three predictors. Mono-syllabicity consistently achieves a significantly higher R^2 value than CVC minimality (Wilcoxon rank sum test, $W = 10^6$, $p < .001$). The chart in (2) shows the same tests on the results from the novel word task experiment in Becker et al. (2008), where Turkish speakers were given stop-final novel nouns of the shapes CVC, CVCC, and CVCVC. Speakers replicated the situation in the lexicon, treating CVCC words as more similar to CVC words than to CVCVC words (Wilcoxon rank sum test, $W = 10^6$, $p < .001$). Neighborhood density is shown to be poorly correlated with alternations in Turkish, as
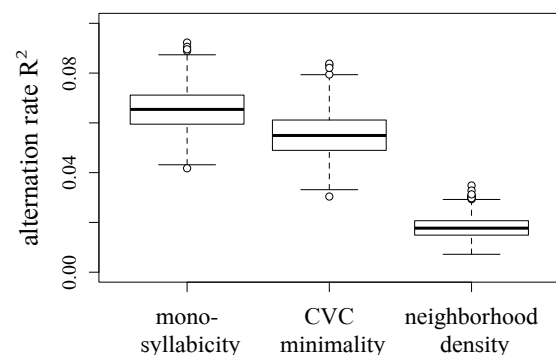
(1)   Turkish lexicon (Inkelas et al. 2000)
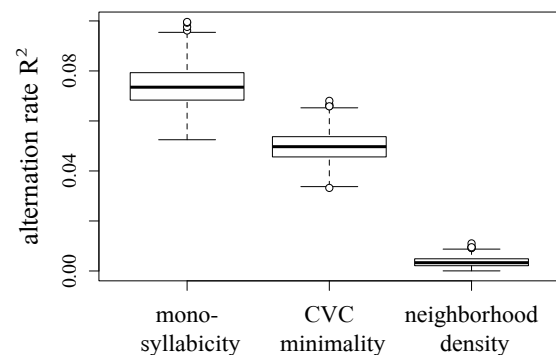
(2)   Turkish novel words (Becker et al. 2008)
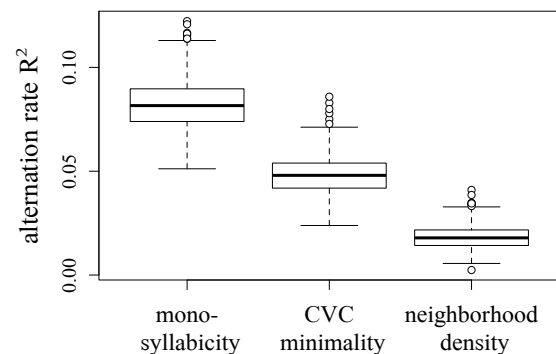
(3)   Russian lexicon (Sharoff 2005)

pointed out by Pycha et al. (2007).

In Russian (3), like in Turkish, final stops exhibit voicing alternations (e.g. $xl^jep \sim xl^jeb$-a 'bread'). The size effect in Russian is the mirror image of Turkish: Stops in shorter words are *more* likely to alternate than longer words (data from Sharoff 2005). Mono-syllabicity is again more tightly correlated with voicing alternations than CVC minimality (Wilcoxon rank sum test, $W = 8*10^5$, $p < .001$). Neighborhood density is, like in Turkish, a poor predictor of alternations. However, because shorter words have more neighbors, and shorter words are more likely to alternate in Russian, neighborhood density is *positively* correlated with alternations in Russian, not negatively as in Turkish, casting serious doubt on Wedel's (2002) idea that dense lexical neighborhoods penalize alternations.

The initial-syllable faithfulness ranking schema IDENT$_{\sigma 1}$ ≫ Markedness ≫ IDENT (Steriade 1994, Beckman 1997, 1998; Casali 1998) was augmented by Becker et al. (2008) with constraint cloning (Pater 2006, 2009). IDENT(voice)$_{\sigma 1}$ has one clone that ranks over the markedness constraint *VTV (which compels alternation of postvocalic voiceless stops when affixed with a vowel-initial suffix), and lists the mono-syllabic stems that don't alternate (in Turkish, 88% of all monosyllables), and the other clone lists the 12% of monosyllables that require low-ranking faithfulness. Poly-syllabic stems are left over for listing by the two clones of IDENT(voice), with only a minority (41%) associated with high-ranking faithfulness vs. 59% with low-ranking faithfulness (4).

(4) Turkish grammar: IDENT(voice)$_{\sigma 1 \; 88\% \text{ of monosyllables}}$, IDENT(voice)$_{41\% \text{ of polysyllables}}$ ≫ *VTV ≫
      IDENT(voice)$_{\sigma 1 \; 12\% \text{ of monosyllables}}$, IDENT(voice)$_{59\% \text{ of polysyllables}}$

Russian has the same constraint ranking as Turkish, but the proportions differ: The proportion of poly-syllables listed with high-ranking faithfulness is higher than it is for mono-syllables. The analysis developed for Turkish, then, extends straightforwardly to Russian, substantiating the breadth of its empirical coverage.

(5) Russian grammar: IDENT(voice)$_{\sigma 1 \; 67\% \text{ of monosyllables}}$, IDENT(voice)$_{89\% \text{ of polysyllables}}$ ≫ *VTV ≫
      IDENT(voice)$_{\sigma 1 \; 33\% \text{ of monosyllables}}$, IDENT(voice)$_{11\% \text{ of polysyllables}}$

In sum, our statistical analysis shows that the word-size based skewed distributions of voicing alternations in Russian and Turkish (and speakers' knowledge of these patterns in novel generalization tasks) are best modeled in formal means by initial-syllable faithfulness, which directly allows for differential treatment of monosyllabic and polysyllabic words. Other analyses (in terms of mora-counting or lexicon-based measures) offer less coverage of the variation in the data, or fail to extend to the distribution of alternations in Russian. Initial-syllable faithfulness, originally motivated largely by static distributional evidence, thus emerges as a multi-purpose formal tool for keeping track within the grammar of distributional asymmetries in alternations as well.

Selected references

Becker, Michael, Nihan Ketrez & Andrew Nevins (2008). The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish devoicing neutralization. ROA-1001.

Ernestus, Miriam & R. Harald Baayen (2003). Predicting the Unpredictable: Interpreting Neutralized Segments in Dutch. *Language* **79**. 5–38.

Pycha, Anne, Sharon Inkelas & Ronald Sprouse (2007). Morphophonemics and the Lexicon: A Case Study from Turkish. In M. J. Solé, P. Beddor & M. Ohala (eds.) *Experimental Approaches to Phonology*, Oxford University Press. 369–385.

Sharoff, Serge (2005). Methods and tools for development of the Russian Reference Corpus. In D. Archer, A. Wilson & P. Rayson (eds.) *Corpus Linguistics Around the World*, Rodopi. 167–180. http://www.comp.leeds.ac.uk/ssharoff/frqlist/frqlist-en.html.

Wedel, Andrew (2002). Phonological alternation, lexical neighborhood density and markedness in processing. Handout from presentation at LabPhon 8, Yale University.