On lexical phonotactics and segmentability

Robert Daland[1], *Benjamin Börschinger*[2], *Abdellah Fourtassi*[3]

[1]Department of Linguistics, UCLA; [2]Department of Computing, Macquarie University; [3]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS, Paris

{r.daland, abdellah.fourtassi}@gmail.com , benjamin.borschinger@mq.au.edu

Word segmentation is the perceptual process(es) by which infant and adult listeners parse speech into a sequence of word-sized units, even in the absence of previous experience with those words. A variety of computational models of word segmentation have been proposed, relying on phonotactics (e.g. Daland & Pierrehumbert, 2011), lexical coding (e.g. Goldwater et al., 2009), or some hybrid approach (e.g Fleck, 2008). Since there are excellent electronic resources for English, most computational models are initially applied to English; dismayingly it has been found that in all cases where a segmentation model is applied to another language, the performance drops relative to English, sometimes drastically so, and regardless of whether the model is more phonotactically or lexically oriented (Daland & Zuraw, 2013; Fleck, 2008; Fourtassi et al., 2013). This fact is dismaying because it is generally assumed that the developmental trajectory of word segmentation does not differ strongly across languages, which implies that the right model of word segmentation should exhibit a similar robustness to cross-linguistic variation; this correspondingly suggests that no extant model is even close to right. However, it is also possible that the modeling results are telling us something important -- that languages really do (or can) vary in their segmentability, as suggested independently by Daland & Zuraw (2013) and Fourtassi et al. (2013). These two papers studied Korean and Japanese, respectively, and proposed that the lower segmentability (relative to English) arose from phonotactic properties, such as the restricted syllable inventory. The present paper explores this line of reasoning further by generating artificial languages in which the phonotactic structure is systematically varied, while controlling other factors that might influence segmentability, such as the word frequency distribution.

Method. Artificial language corpora were generated in three stages: defining a *grammar*, generating a *lexicon/corpus* according to a frequency distribution, and then applying an unsupervised word *segmentation* algorithm. *Grammars*. The alphabet ∑ consists of a 'typical' segmental inventory of 13 consonants [ptkfsxmnŋlɾwj] and 5 vowels [aeiou]. Grammars were stated in the Maximum Entropy Harmonic Grammar formalism, in which constraints are stated as illicit sequences of natural classes using SPE-style featural descriptors, e.g.the constraint banning all consonant clusters is written '*[-syl][-syl]'. The 'harmony' of a string in ∑* is the weighted sum of its constraint violation, and the probability of a string is proportional to the exponential of its harmony (Hayes & Wilson, 2008). The maximally restricted language, **Pseudo-Senufo**, is defined as a strict CV language (words cannot begin with V; words cannot end with C; no CC; no VV). The maximally unrestricted language, **Pseudo-Berber**, includes no phonotactic constraints at all (so that the form [tfktkx] would be just as probable as the form [patiku]). An intermediate language, **Pseudo-Korean**, includes phonotactic properties reminiscent of Korean (maximally CGVC syllables, coda neutralization, [l]/[ɾ] allophony, Syllable Contact Law). *Lexicons/corpora*. Lexicons were sampled from the probability distribution over ∑* assigned by the grammar, subject to a maximum word length of 6 and a modest penalty on every segment (*Struct) which served as a soft preference for shorter words. Lexical types were then assigned frequencies, and tokens were scambled, so as to match the type-token and utterance-length distributions in a subset of the Brent-Bernstein-Ratner corpus. *Segmentation*. The Adaptor Grammar formalism assigns a posterior distribution over segmentations of the corpus, according to a generative model of the language (Johnson et al., 2007). Sample segmentations are generated according to a Markov Chain Monte Carlo process, and the final segmentation is generated by applying Minimal Bayes Risk decoding to posterior samples. The token f-scores are as follows: Pseudo-Berber 93%; Pseudo-Korean 92%; Pseudo-Senufo 83%.

Discussion. The results suggest that when distributional factors are controlled, languages which are more phonotactically restricted are predicted to be harder to segment, consistent with previous proposals (Daland & Zuraw, 2013; Fourtassi et al., 2013). At present, it is unclear whether every phonotactic constraint is always predicted to reduce segmentability; for example, differential rates of assimilation between and across word boundaries might make some phonotactic cues stronger (Daland & Pierrehumbert, 2011). This abstract reports on 3 grammars; the final results are expected to include additional languages, such as a variant of Pseudo-Korean with the Syllable Contact Law constraints removed, and variants of Pseudo-English with stronger and weaker degrees of Sonority Sequencing Principle for onsets.

References. Daland, R. & Pierrehumbert, J.B. 2011. Learning diphone-based segmentation. *Cog. Sci.*, 35(1), 119-155. / Daland, R. & Zuraw, K. 2013. Does Korean defeat phonotactic word segmentation? *Proc. 51st ACL, Sofia, Bulgaria*. / Fleck, M.M. 2008. Lexicalized phonotactic word-segmentation. *Proc. 45th ACL: HLT*, 130-138. / Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. 2013. Whyisenglishsoeasytosegment? *Proc. 51st ACL, Sofia, Bulgaria*. / Goldwater, S., Griffiths, T.L. & Johnson, M. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112, 21-54. / Johnson, M., Griffiths, T.L., & Goldwater, S. 2007. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. *NIPS* 19: MIT.