Gradient Word-Edge Statistics Influence Syllable Segmentation Judgements

Previous work modeling syllable boundary learning and phonotactic learning ([1][2]) makes the simplifying assumption that target syllabifications are defined using the Maximum Onset Principle(MOP), with boundaries placed categorically to yield the longest possible legal syllable onsets [3]. However, recent empirical data on syllable boundary locations supports variation and additional influences alongside MOP ([4,5,6,7,5]). Additionally, much research has found that people are sensitive to the statistical distributions of sound categories in learning language-specific word segmentation (e.g.[8]). Given segmentation's sensitivity to statistics and uncertainty about the learning and representation of word-medial syllable boundaries, I use model comparison to evaluate whether the gradient, joint lexical frequencies of word-initial and word-final consonant sequences contribute to English speakers' syllabifications, above and beyond other factors such as MOP.

The relative frequency of each consonant sequence (from the vowel to the word edge) was estimated word-initially and word-finally from the Carnegie Mellon pronunciation dictionary . For the possible syllabifications of a word, I calculated its "JointWordEdgeScore": the product of the frequencies of the resulting onset and coda (e.g. in 'capstan' [kæpstən], the JointWordEdgeScore for [kæps.tən] is the product of the word-final frequency of [ps] and the word-initial frequency of [t], and the score for [kæp.stən] is the product of the word-final frequency of [p] and the word-initial frequency of [st]).

A logistic regression (11) model was fit to [Eddington]'s experimental English syllabification judgements. One model's predictors included known English syllabification factors (MOP, vowel stress) alongside normalized JointWordEdgeScores, and another model was the same except that it lacked the JointWordEdgeScores. Table 1 shows that the inclusion of the JointWordEdgeScore predictor improved model BIC, corresponding to a Bayes factor of >100, indicating decisive evidence for the model with JointWordEdgeScore over the model without it ([12]).

This model comparison suggests that gradient word edge statistics English word-medial syllabification, with potential implications for syllable representations and learning; just as [13] proposes utterance boundary statistics bootstrap word segmentation, word boundaries might bootstrap syllable segmentation gradiently.

References [1] Goldwater et al. (2005). Representational bias in unsupervised learning of syllable structure. In Proceedings of computational natural language learning (conll-2005). [2] Daland, R. et al. (2011, Aug). Explaining sonority projection effects. Phonology. [3] Kahn, D. (1976).Syllable-based generalizations in English phonology. [4] Hong, S.-H. (2021). A weighted constraint grammar analysis of word-medial syllabification in English. Linguistic Research. [5] Sturm, P. (2018). Experimental evidence on the syllabification of two-consonant clusters in Czech. Journal of Phonetics [6] Eddington,D. et al. (2013).Syllabification of American English Evidence from a large-scale experiment. Journal of Quantitative Linguistics [7] Berg, Thomas et al.. "Syllabification in Finnish and German: Onset Filling vs. Onset Maximization." *JP* 2000 [8] Rubach, Jerzy, and Geert Booij. "Syllable Structure Assignment in Polish." *Phonology* 1990 [8] Thiessen, Erik D., and Lucy C. Erickson. "Beyond Word Segmentation" *Current Directions in Psychological Science* 2013 [9] Saffran, Jenny R., et al. . "Pattern Induction by Infant Language Learners." *Developmental Psychology* 2003. [11] Mayer, C. et al. maxent. ot a package for doing maximum entropy optimality theory (2022) [12] Kass, Robert E., and Adrian E. Raftery. "Bayes factors." *Journal of the American Statistical Association* (1995) [13] Batchelder, Eleanor Olds. "Bootstrappincomputational model of infant speech segmentation." *Cognition* (2002)

Model Predictors	**Log Likelihood of Experimental Data**	**BIC**	**ΔBIC**
Morph, Stress, OnsetMax	-51133.3	102300.9	1525.1
Morph, Stress, OnsetMax, JointWordEdgeScore	-50365.01	100775.8	0

Table 1. Model comparison of MaxEnt (equivalent to multinomial logistic regression) models with and without normalized joint word edge statistics as a predictor. Morph is a predictor that favors syllabifications that align with a morpheme boundary, Stress favors syllabifications where stressed syllables receive codas, and OnsetMax favors syllabifications that follow MOP.