Gradient Word-Edge Statistics Influence Syllable Segmentation Judgements

Cerys Hughes
ceryshughes@umass.edu
UMass Amherst
HILS 2024

Motivation

- How are word-medial syllable boundaries learned?
- Simplifying assumption in previous work: boundaries from by Maximum Onset Principle.

Empirical data: variation and additional influences.

Word segmentation: sensitivity to statistics

Is syllabification sensitive to gradient, joint lexical statistics from word edges?

Maximum Onset Principle(MOP)

- Word-medial syllable boundary placement to yield longest possible legal onset (Kahn 1976)
- kaep.stən, not kaeps.tən

Where are syllable boundaries?

- Variable and non-MOP medial syllable boundary judgements:
  - Polish (Rubach & Booij 1990) , French (Content et al. 2001), Irish (Chiosain et al. 2012), Czech(Šturm 2018), English (Eddington et al. 2013)
- Eddington et al. (2013):
  - ~5,000 English lexical items
  - Each presented to ~25 participants
  - Forced choice syllabification judgement

Joint Word Edge Score

- Frequency estimates from CMU Pronunciation Dictionary

P(Onset): estimated from word-initial

P(Coda): estimated from word-final

JointWordEdgeScore of a syllabification: P(Onset) × P(Coda)

JointWordEdgeScore of kaep.stən = P(Onset = st) × P(Coda = p)

- Normalized for each lexical item
- Accounts for coda

Hypothesis

- Assuming a logistic regression model for further analyzing Eddington et al. (2013)'s experimental data.
- Including known predictors of syllabification
- If English speakers' syllable boundary representations are sensitive to joint word-edge statistics, then:
  - Including the joint word-edge statistics predictor will significantly improve model fit

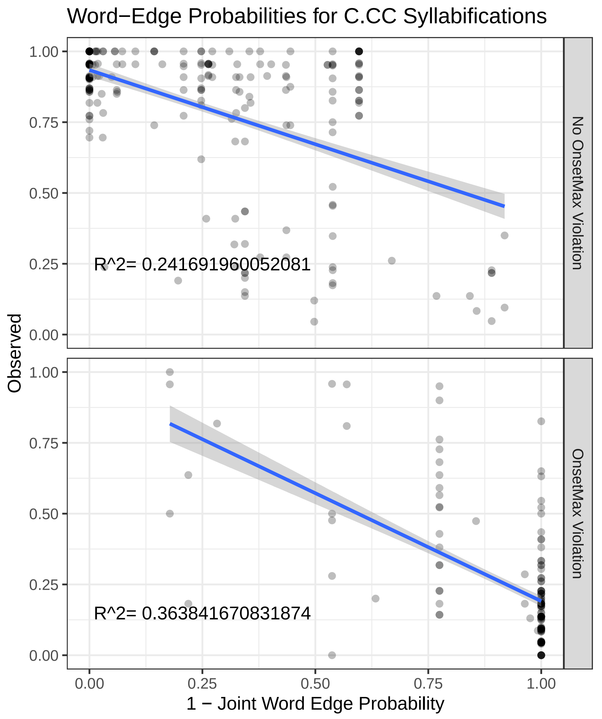
Results: Model Comparison

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

Table: Model comparison of multinomial regression with and without the inclusion of word-edge statistics (the JointWordEdgeScore predictor).

Bayes Factor of > 100: decisive evidence for JointWordEdgeScore model

Results: Joint Word Edge Score Distribution



- Each point: the C.CC syllabification option for one lexical item (with 3 medial consonants).
- Vertical: how often participants chose this syllabification response. Horizontal: normalized Joint Word Edge Score of the candidate.
- Top: candidates obeying MOP

Discussion

- Supports the hypothesis that joint word-initial and word-final statistics influence English word-medial syllabification
- Similar results found for word-initial statistics in Czech: Šturm (2018)
- Possible implications for syllable representations and learning:
  - Do statistics at word boundaries bootstrap learning of word-medial syllable boundaries?
  - Are some word-medial syllable boundaries gradient?
- More implicit measures of speakers' word-medial syllable boundaries?
- Correlation between MOP and lexical statistics?
- Incorporating joint word edge scores in models of syllable segmentation learning?

Conclusion

- Simple measure of scoring syllabification based on word-initial and word-final lexical statistics
- Model comparison shows English speakers' syllabifications partly explained by joint onset, coda lexical statistics, beyond MOP

References

[1] Goldwater et al. (2005). Representational bias in unsupervised learning of syllable structure. In Proceedings of computational natural language learning. [2] Daland et al. (2011). 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 et al. (2013).Syllabification of American English Evidence from a large-scale experiment. Journal of Quantitative Linguistics [7] Berg et al. "Syllabification in Finnish and German: Onset Filling vs. Onset Maximization." JP 2000 [8] Rubach, J., and Booij G. (1990) "Syllable Structure Assignment in Polish." Phonology 1990 [8] Saffran et al. "Pattern Induction by Infant Language Learners." Developmental Psychology 2003. [9] Mayer et al. maxent. ot a package for doing maximum entropy optimality theory (2022) [10] Kass, R., and Raftery A. "Bayes factors." Journal of the American Statistical Association (1995) [11] Batchelder, E. "Bootstrapping computational model of infant speech segmentation." Cognition (2002)