

A large-scale corpus investigation of Italian Troncamento

Breiss Lab (Dhanya Charan (CMU), Yue Chen (USC), Darby Grachek (USC), Po-Hsuan Huang (USC), Muxuan He (USC), Haley Hsu (USC), Jane Li (Johns Hopkins U.), Yang Wang (U. Utah), Canaan Breiss (USC)) - lab.cbreiss.com

1. Background

Troncamento: Word-final mid-vowels [e, o] preceded by sonorants are deleted between words: /e, o/ → ∅ / [+son] _ {C, V}

- (1) di colore verde scuro → di color_ verde scuro
of color green dark “Of dark green color”

Troncamento is *subject to variation* even when its phonological environment is satisfied (obligatory in some syntactic environments and banned in others)

Big-picture question: How do grammatical and processing constraints *interact* in Troncamento?

Today’s question: What grammatical and processing factors condition the variation in Troncamento?

2. Structural / Grammatical factors

Previous analyses rely largely on **syntactic structure** to account for varied application (Meinschaefer, 2005; Vogel et al., 1983)

Obligatory	Banned
(1) Infinitive + enclitic pronoun sequences	(1) Syntactic environments: between subject DP + VP, post-nominal adjectives, noun + verb, noun + a PP that it doesn’t c-command
(2) Modal verb + infinitive	(2) Some specific lexical items
(3) Nouns as titles + bare noun sequences	(3) Between pauses (Nespor, 1990; Tanner et al., 2017)

3. Processing / Cognitive factors

Wagner (2012) proposes locality of production planning can explain observed morphosyntactic locality effects:

Production Planning Hypothesis (PPH): Phonological processes apply blindly across word-boundaries, but phonological environment can only contain following word if has been planned.

We predict that, in addition to structural/grammatical factors, production planning can also constrain Troncamento application:

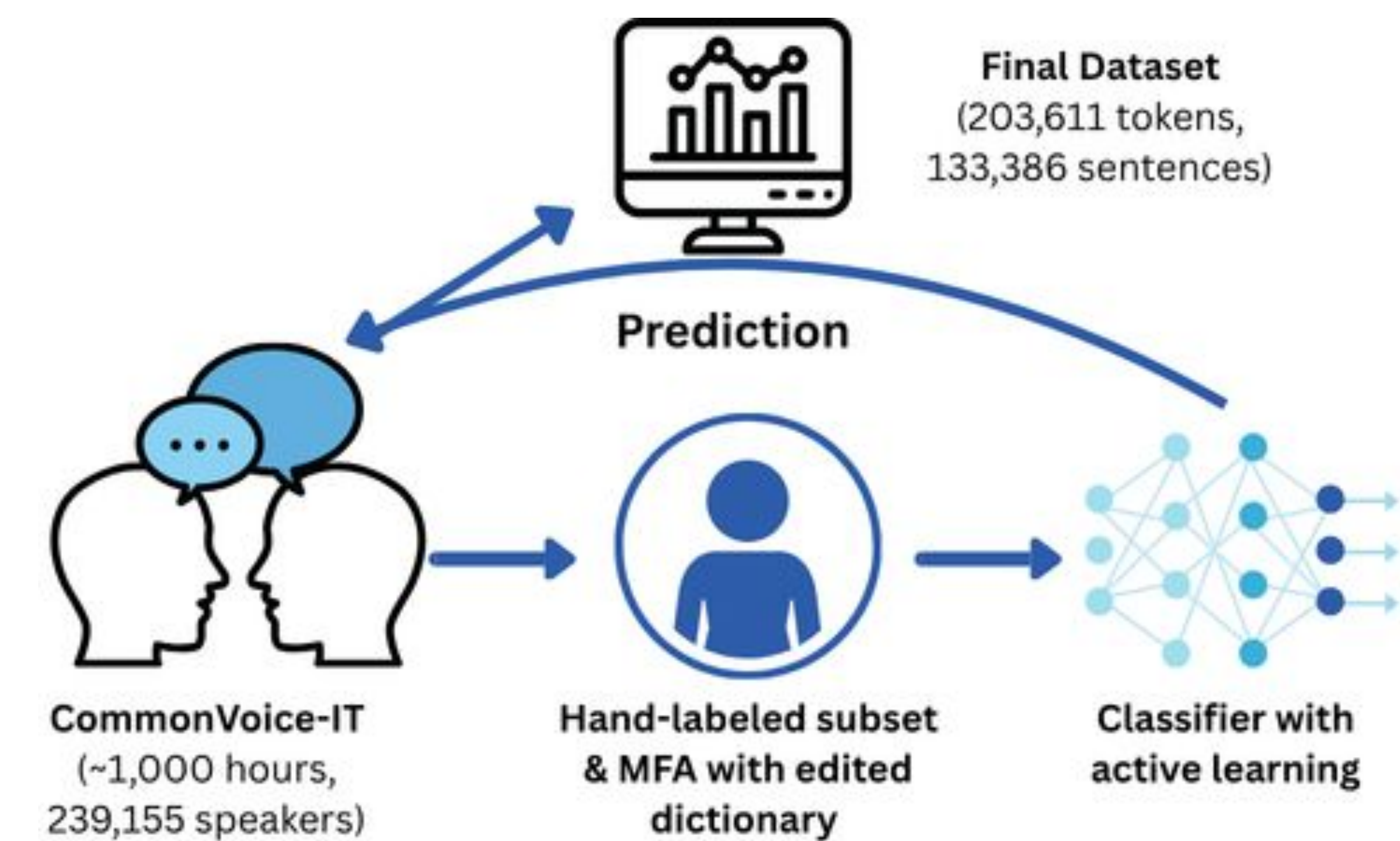
Production planning shown to be approximated by several factors:

- Local speech rate:** n_phones in target word/duration of the word (sec)
Prediction: Faster speech rate = more Troncamento
- Bigram frequency:** log count of the target word + the following word
Prediction: Higher bigram frequency = more Troncamento
- Suprisal:** $P(\text{target word} \mid \text{preceding context in sentence}) / P(\text{following word} \mid \text{preceding context in sentence})$
Prediction: Lower surprisal = more Troncamento

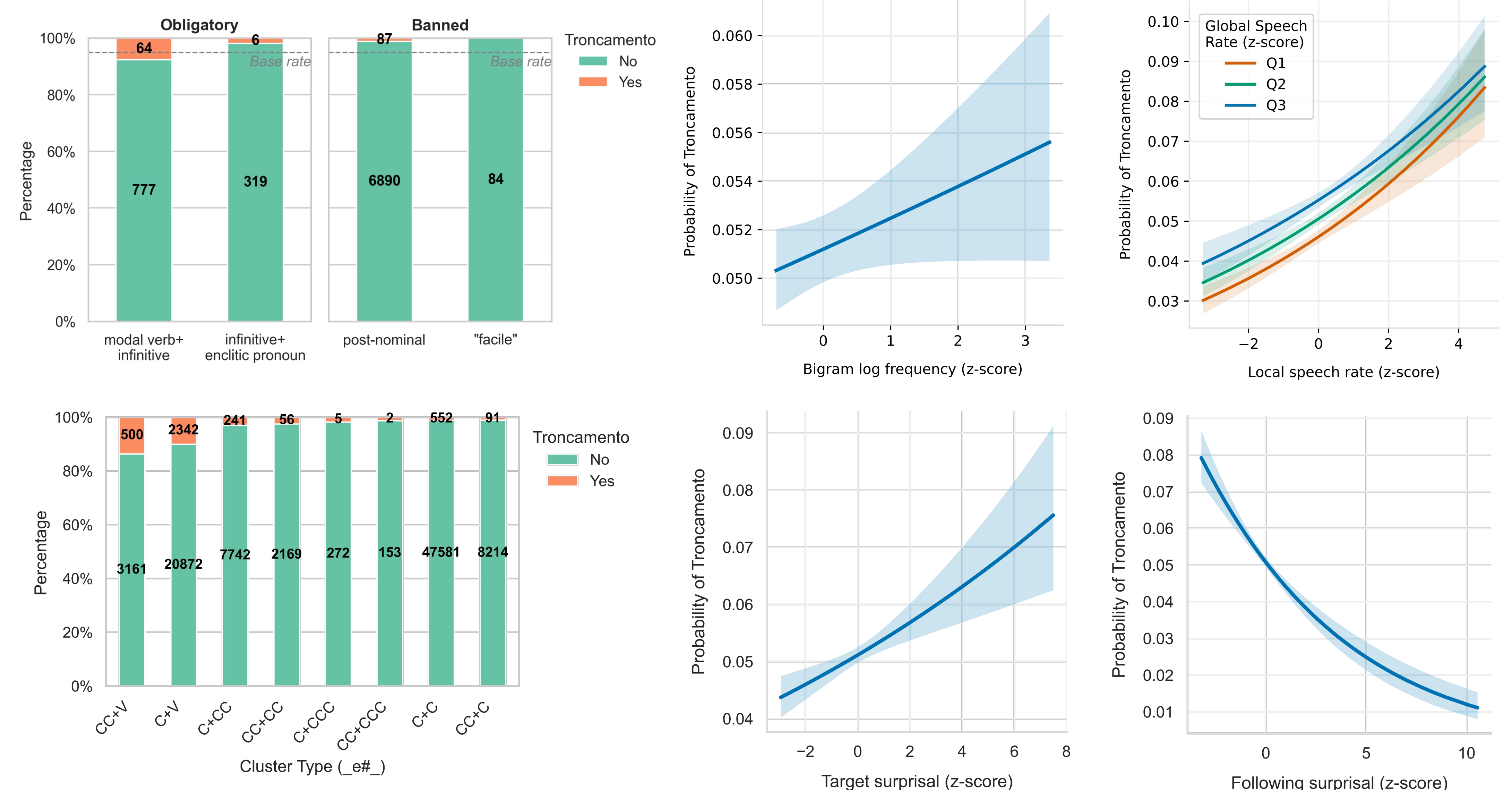
4. Materials and Methods

Active learning loop (Bonwell & Eison, 1991):

1. Train a Troncamento classifier on hand-labeled data
2. Get the most useful unlabeled samples for the classifier
3. Hand-label the samples
4. Retrain the classifier
5. Repeat steps 2-4 until classifier is sufficiently accurate (F-score on validation set > 0.9).



5. Results



6. Discussion

Structural: Mixed evidence for the literature:

- Only *obliquely* related to reported obligatory/banned syntactic contexts
- novel evidence of phonological optimization!

Processing: Clear replication of literature:

- bigram frequency, speech rate, surprisal
- Next up:** do the two classes of factors interact? If so, how?