

Models of Speaker Choice in Production: Integrating Information and Availability

Recent work proposes an information-theoretical account of sentence production, **Uniform Information Density [UID]**; Jaeger06, LevyJaeger06]: speakers structure their utterances so as to maintain a constant rate of information transmission (where $\text{INFORMATION}(\text{word}) = -\log \text{PROBABILITY}(\text{word} \mid \text{context})$). Initial evidence that speakers prefer UID comes from probability-conditioned pronunciation weakening, syllable/word shortening, and optional *that*-insertion. Optional *that* in (a) is more likely with less predictable complement clauses/clause onsets (*we*). An alternative account, **Availability-based Sentence Production [ASP]**; BockWarren85, FerreiraDell00], claims that speakers insert *that* when the complement clause onset is not yet available for pronunciation. We present evidence for the contributions of ASP and UID from three large-scale studies of morphosyntactic contractions in spontaneous speech (b-d).

- (a) You think [(*that*) we should do this]?
- (b) You're / *are* free to leave any time.
- (c) He couldn't / *not* do it.
- (d) They've / *have* been busy lately.

For **Study 1**, we extracted 6,488 reduced and 2,845 full forms of "be" (*BEs*) (b) from a corpus of spontaneous speech. UID predicts that *BE* is more likely to be contracted where it is more predictable (and hence carries less information). In order to compare the predictions of UID and ASP directly, we introduce a predictability-based implementation of ASP. In this formalization, ASP holds that *BE* should be reduced when the word following *BE* is predictable or frequent (*i.e.*, these factors are taken to be measures of accessibility). *BE* predictability and the predictability of the following word were both estimated as probabilities conditioned on the surrounding words. From example (b),

Predictability of BE (UID):

- (1) $P(\text{ARE} \mid \text{you}) = P(\text{'re} \mid \text{you}) + P(\text{are} \mid \text{you})$
- (2) $P(\text{ARE} \mid \text{free}) = P(\text{'re} \mid \text{free}) + P(\text{are} \mid \text{free})$

Predictability of word following BE (ASP):

- (3) $P(\text{free} \mid \text{ARE}) = P(\text{free} \mid \text{'re}) + P(\text{free} \mid \text{are})$

We tested both hypotheses in a series of logit mixed effects model with important additional controls (e.g. speech rate, syntactic context, frequency of surrounding words, etc.). In line with UID, *BE* predictability affects speakers' choice between full and contracted *BE* more than any other factor (factor-removal: $\chi^2(1) = 768.5$, $p < 0.0001$): predictable *BE* is reduced more often. The predictability of the following word, however, has a significant independent effect ($\chi^2(1) = 78.9$, $p < 0.0001$). Consistent with ASP, *BE* is reduced more often before predictable words.

Studies 2 and 3 extend these findings to 5,083 reducible cases of *NOT* (c) and 2,425 cases of reducible *HAVE* (d). In both cases the predictability of the reducible element is the strongest predictor of contraction, with predictability of the upcoming word exerting a smaller (but still significant) effect.

This work shows that UID affects morphosyntactic choice, adding to previously documented effects on choices at other levels of linguistic representation. UID captures speakers' preference to reduce forms that convey redundant information. The importance of the predictability of the word being produced given its context shows that ASP alone is insufficient to explain speakers' choices. We lay out possibilities for combining UID and ASP into a single formal information-based model.