Speakers Embody Fine-Grained Probabilistic Knowledge

Corpus studies of spontaneous speech suggest that grammaticality is gradient (Wasow 2008), contra a widespread assumption. In experiments with the English dative, Bresnan (2007) compared a corpus-derived predictive model to a regression model built from subject trial data, showing that the same probabilistic factors that influence production also influence judgment and are thus arguably part of competence. Her model, however, is itself largely predicated on categorical predictors. In contrast, predictive models of English complement clause (CC) and non-subject-extracted relative clause (RC) 'that'-optionality (Jaeger 2006) have as their most significant factors the predictability of an embedded clause, given the verb (CC) or head noun (RC) lemma—factors that are themselves highly gradient. Establishing that these factors are similarly involved in judgments could provide evidence that fine-grained probabilistic knowledge is part of linguistic competence.

We undertook eight such judgment experiments, employing both existing methods and experiments run via Amazon’s online ‘Mechanical Turk’ platform, providing access to a large pool of participants.

**Experiment 1** (26 participants, 30 items) began with the models of RC-reduction developed in Jaeger (2006). Corpus tokens were then binned by relative model-predicted probability of ‘that’-omission. Six tokens were extracted at random from each of five bins (0≤ <20% likelihood of ‘that’-inclusion; 20≤ <40%; and so on.) In a gradient scoring paradigm—distributing 100 points between available options (Bresnan 2007)—participants rated how likely each choice—with or without ‘that’—was as the continuation of a segment of discourse. As hypothesized, mean participant ratings significantly correlate with corpus model predictions ($r=0.614$, $p=0.0003$).

To address concerns, however, that subjects might be ‘over-thinking’ the process, **Experiment 2** (29 participants) replicated Experiment 1 using a timed forced-choice paradigm where participants had from 5 to 24 seconds (varied as a linear function of token length) to choose between the reduced/unreduced RC stimuli. These results correlate even more closely with predictions ($p=0.838$, $p<0.0001$).

Seeking both to generalize across populations and to explore a methodology of potential to psycholinguistics, **Experiments 3 and 4** replicated RC Experiments 1 and 2 on Amazon’s Mechanical Turk (1600 hits each, across 89 and 66 unique participants, respectively). While Turk measures are expected to be noisier (lower correlations) due to unbalanced data sets, the results remain significant in both trials ($r=0.562$, $p=0.0009$; $r=0.364$, $p=0.0285$), offering evidence of the robustness of the alignment between production and judgment models.

**Experiment 5-8** employed the same paradigms (100-point and timed forced-choice, lab-based survey and via Turk) to investigate ‘that’-mentioning in CCs, where predictability of embedding is an even stronger factor in the corpus model. As with the first four experiments, each of these trials produced significant correlations ($r=0.433$ to 0.500, $p=0.0107$ to 0.0034). Finally, mixed-effect logit models were fitted to the judgment data. Exactly as in the corpus-derived models, predictability of embedding remains the most significant factor in all experimental models (excluding individual subject variation).

In conclusion, these results suggest that speakers consider the same factors in judgment as in production, offering evidence that competence grammar includes access to probability distributions.