Speakers Employ Fine-Grained Probabilistic Knowledge

Tests of readers’ judgments regarding optional ‘that’ introducing relative clauses and complement clauses provide evidence that fine-grained probabilistic knowledge is involved in such judgments and is arguably part of linguistic competence. Participants judged the relative naturalness of pairs of sentences, identical except for the presence or absence of ‘that’, as in examples (1) and (2).

(1) **Complement clause:** So you understand (that)\textsubscript{COMP} this was a long time ago.
(2) **Relative clause:** I have a few favorites (that)\textsubscript{REL} I use more than others.

The results reveal a strong correlation between a corpus-based model’s predictions regarding ‘that’ occurrence and participants’ judgments of where ‘that’ should be used.

Both clause types were tested in four ways, using two methods of comparison and two populations of participants. One method had participants distribute 100 points between the versions with and without ‘that’; the other was a binary forced-choice, under time constraints. Both methods were employed in a traditional lab setting and also crowd-sourced via Amazon’s Mechanical Turk (AMT) facility (Munro et al., 2010).

In all cases, the stimuli were sentences taken from the Switchboard Corpus (Godfrey et al., 1992), presented with some preceding context. We began with Jaeger’s (2006) existing corpus models of ‘that’ distribution. Tokens were binned by model-predicted probability of ‘that’-omission. Six tokens were extracted at random from each of five bins (0≤p<20% likelihood of ‘that’-inclusion; 20≤p<40%; and so on).

All of the relative clause studies yielded significant correlations for participant judgments and corpus model probabilities, with coefficients ranging from .365 to .838. The complement clause studies likewise all yielded significant correlations, with correlation coefficients ranging from .433 to .500. AMT measures are expected to be noisier (lower correlations) due to unbalanced data sets, but the fact that they were significant here and agreed with the laboratory studies suggests that web-based crowd-sourcing offers a viable alternative to traditional lab methods.

Finally, mixed-effect logit models were fitted to the judgment data. As in the corpus-derived models, the same factor—predictability of embedding—remains most significant in all experimental models (excluding individual participant variation), offering further evidence of the robust alignment between production and judgment.

These results suggest that competence grammar includes not only some degree of probabilistic information, as suggested by prior work, but even access to rather fine-grained probability distributions. Previous complementary judgment experiments (Bresnan, 2007) argue that probabilistic factors involved in production also influence judgment, but the models in those studies include largely categorical predictors. With English complement clause (CC) and relative clause (RC) ‘that’-optionality, however, the predictive models shown here to closely align with both judgment and production data turn most significantly on the predictability of an embedded clause, given verb (CC) or head-noun (RC)—measures that are highly gradient.
REFERENCES


Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen and Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, Los Angeles.