

Combining web counts and lexical semantics to predict that-omissibility

September 5, 2008

The optional complementizer *that* has captured the interest of linguists and psycholinguists alike. According to Thompson & Mulac's (1991) classic analysis, *that*-omission is likely after epistemic uses of verbs like *think* and *guess*. Most subsequent analyses have focused on factors such as verb frequency and the planning difficulty of the subordinate clause following *that* (e.g. Jaeger 2006, Ferreira & Dell 2000). For example, *that*-preference correlates negatively with verb frequency as well as with SC-subcategorization bias, i.e. the probability with which the verb takes a sentential complement. We argue that the resulting models, which lack semantic factors, can be improved by taking into account an extension of Thompson & Mulac's analysis. We argue that the semantic property promoting *that*-omission is not just epistemicity of the matrix verb, but semantic bleaching in general. This allows us to account for the behavior of verbs conveying mood or affect, such as *wish*, *regret* and *hope*, which have low *that*-preference, but don't convey epistemicity.

We developed a rating task for obtaining subjective measures of the semantic lightness associated with verbs of varying degrees of *that*-preference. Subjects were asked to rate the mental effort ascribed to a hypothetical speaker of randomly constructed sentences such as "I know that hypnales have human souls," or "I assume that leucrota can do sums." We were also interested in the feasibility of using Google-counts as corpus data. We therefore constructed a linear regression model using measures of *that*-preference, verb frequency, and SC-bias based on Google counts for the strings $I \langle verb \rangle$, $I \langle verb \rangle he$, and $I \langle verb \rangle that he$. These contexts were chosen as a means of (a) excluding noun uses of forms like *guess*, (b) locating constructions that are unambiguously biclausal, and (c) including epistemic and mood/affective discourse uses of verbs. The Google counts significantly but weakly correlate with previously collected norming data on *that*-preference and SC-bias (Garnsey et al., 1997; $F(1, 46) = 16.5, r^2 = .26, p < .001$ for *that*-preference; $F(1, 46) = 10.4, r^2 = .18, p < .005$ for SC-bias).

We then added the subjective ratings as a predictor to a linear regression model with *that*-preference as the outcome variable and previously available data on verb frequency and SC-bias as predictors. The subjective ratings significantly improved model prediction, as revealed by ANOVAs comparing models with and without the subjective ratings ($F(1, 36) = 16.2, p < .001$).

Our interpretation of the subjective ratings is that any factor reducing processing effort, such as high frequency, predictable subcategorization, and semantic bleaching of the matrix verb, promotes *that*-omission. Our study adds to a growing literature utilizing web counts, which opens up exciting — and time-saving — prospects for further psycholinguistic research.

References

- Ferreira, V. S. and G. S. Dell, 2000. 'The effect of ambiguity and lexical availability on syntactic and lexical production'. *Cognitive Psychology* **40**, 296–340.
- Garnsey, S. M., N. J. Perlmutter, E. Meyers, and M. A. Lotocky, 1997. 'The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences'. *Journal of Memory and Language* **37**, 58–93.

- Jaeger, T. Florian, 2006. Probabilistic Syntactic Production: Expectedness and Syntactic Reduction in Spontaneous Speech, Linguistics Department, Stanford University.
- Thompson, S. A. and A. Mulac, 1991. 'The discourse conditions for the use of complementizer *that* in conversational English'. *Journal of pragmatics* **15**, 237–251.