Judgment Studies of Information-Driven Syntactic Reduction

Robin Melnick
rmelnick@stanford.edu

1. Introduction

While corpus-based language studies are perhaps the most common medium for sociolinguistic research, they are comparatively rare in psycholinguistic work (though cf. Clark & Fox Tree 2002; Szmrecsanyi 2005; inter alia). Corroborating experimental results, however, can substantially bolster initially corpus-supported hypotheses. For example, evidence from judgment experiments (Bresnan 2007; Bresnan & Ford 2010) has complemented and further validated a previously developed, corpus-derived model for probabilistic production of the English dative alternation (Bresnan et al. 2007). Having then shown that the same factors appear to underlie both production (corpus) and judgment (experiment), it is argued that this alignment of complementary evidence suggests that speakers’ implicit knowledge of their language fundamentally includes such probabilistic information (Bresnan 2007:17; Bresnan & Ford 2010:189).

In this light, the main purpose of the present study is to provide similar experimental support for a hypothesis of syntactic reduction as being substantially influenced by information-theoretic considerations, an argument previously made on the basis of corpus (production) data (Jaeger 2006, 2010).

1.1. Information Theory in Linguistic Analysis

Any discussion of information theory begins with Shannon’s seminal proposal that communication, most generally stated, is maximally efficient when peaks and troughs in information density are avoided in the flow of data through a noisy channel (Shannon 1948). Per Shannon, information content is inversely proportional to contextual probability. Applied to language, the more likely—or expected—a word, syntactic construction, or any other linguistic element is in its context, the less informative it is in such context. Put another way, the more probable an element, the more “redundant” (Jaeger 2006) for communicative purposes.

Thanks first to my qualifying paper review committee: Joan Bresnan, Martin Kay, and Tom Wasow. Special thanks to Florian Jaeger for sharing his databases of complement- and relative-clause constructions developed in his work on syntactic reduction (Jaeger 2006, 2010). Thanks also to attendees of the Stanford Psychology of Language Tea (SPLaT) and Spoken Syntax Group for their insightful comments on preliminary versions of this material. All graphics and statistical calculations were prepared using R software version 2.12.1 (R Development Core Team 2010).
In turn, we can view information theory applied to linguistic analysis as building on the earlier observation (Zipf 1935) that frequent words in a language tend to have shorter forms. The extension from Zipf made by information theory, however, is that not only should we find that frequent elements are correlated with shorter forms in the lexicon, we should also find that such forms are further reduced in production when they are most likely given their production context. Aylett & Turk 2004 find exactly that: Independent of form in the lexicon, words tend to be further phonetically reduced when most predictable. Bell et al. (2003) look at such probability-sensitive production in coordination with other processing pressures—disfluencies, speech rate, and so on—and model the several effects via multivariate regression. Finally, Bell et al. (2009) extend this line of research in finding that while both content and function words are affected by such conditional contextual probability, only short functional words appear to be affected by *preceding* context.

Beyond application of information theory at the word level, Tily et al. (2009) found an interaction of syntactic probability and phonetic production in English datives. The likelihood from context of prepositional or double-object dative alternatives was seen to significantly affect the observed duration of *to* in the prepositional construction, as well as the rate of disfluencies in both constructions. In comprehension, as well, Jurafsky (1996) observes listener sensitivity to probability distributions in syntactic disambiguation, and in extending this, Hale (2003) found that well-known grammars for phenomena such as garden-pathing, center-embedding, and subject-object extraction asymmetry are predicted by probabilistic grammars with sensitivity to information density. Above the level of syntax even, Genzel and Charniak (2002) noted an information-theoretically efficient distribution of information both through the production of a given sentence and across multi-sentential discourse.

In turn, Levy and Jaeger (2007) have suggested that in production, the predictability in context of linguistic elements affects not only phonetic form or measures of fluency but a speaker’s online choice between syntactic alternatives. Combining this with the predictions of the several other aforementioned works on information-theoretic linguistic variation, the authors label this collective notion *Uniform Information Density* (UID) and predict that we should observe its effects at all levels of language production (Levy & Jaeger 2007; Levy 2008; Jaeger 2010). Specifically, UID posits that, within bounds allowed by a language’s grammar, speakers act where possible—i.e., at choice points within the production of natural speech—to spread out the information content contained in an utterance so as to keep the density, or flow, of such information at, or as close as possible to, a constant rate.

1.2. **UID case in point: optional ‘that’**

Where the Jaeger (2010) argument for UID is supported by corpus-based modeling of factors underlying syntactic reduction, the present work seeks evidence of like information-theoretic considerations in experimental judgments of the same
syntactic reduction phenomenon: optional *that*-mentioning with English relative and complement clauses. Specifically, the prior corpus-based work here (Wasow et al., in press; Jaeger 2006, 2010) suggests that when the onset of a relative clause (RC) or complement clause (CC) is comparatively surprising in context—carries high Shannon information load—speakers are more likely to act to spread out this information by including the optional complementizer *that*. Where RC or CC onset is relatively unsurprising in context, Jaeger’s corpus-derived model predicts that speakers are comparatively less likely to include optional *that*.

Since estimating the predictability of such a clause conditioned on its complete context—including syntax, prosody, even the entirety of a speaker’s encyclopedic knowledge—is infeasible, prior works (Jaeger 2006, 2010) and the current study alike rely on a rather simple approximation: for an RC, just the embedding relativized nominal head (lemma); for a CC, just the embedding matrix verb (lemma). The likelihood of CC, for example, given context is then estimated as the conditional probability of the onset of a complement clause given matrix verb, as indicated in (1):

\[
p(\text{CC}|\text{context}) \cong p(\text{CC}|\text{matrix verb lemma})
\]

The information content \(I\) represented by such a CC onset is thus:

\[
I(\text{CC}|\text{context}) \cong -\log p(\text{CC}|\text{matrix verb lemma})
\]

It is predicted that where the probability of a CC is greater—and thus its information content less—given a matrix verb’s frequency of embedding relative to its overall frequency in the language, participants will judge optional *that* as being less likely, matching the predictions of Jaeger’s extant corpus model.

For example, in the Switchboard corpus of spontaneous speech (Godfrey et al. 1992), on which the Jaeger CC *that*-mentioning model is based, a much higher proportion of \([v \text{ think}]\) tokens take a complement clause than is the case for \([v \text{ confirm}]\).

\[
\begin{align*}
(3)\ a. \ & \text{My boss } \text{thinks } [ (\text{that}) \text{ we are absolutely crazy }]. \\
& \text{My boss } \text{confirmed } [ (\text{that}) \text{ we are absolutely crazy }].
\end{align*}
\]

The information content associated with the onset of CC in (3a) with matrix verb lemma *think* is therefore less than that associated with the CC in (3b) with matrix verb lemma *confirm*. Optional *that*-mentioning is then predicted to be comparatively less likely with (3a) than (3b). (Examples from Figure 1, Jaeger 2010:27)

1.3. Alternative Accounts

Certainly, UID is far from the only explanation put forward to account for choices in syntactic variation, including optional *that*-mentioning among other alternations.
A wide range of prior psycholinguistic work has explored potential processing constraints on acquisition, production and comprehension, including availability/accessibility (Bock & Warren 1985; Bock 1986; Ferreira & Dell 2000; Race & MacDonald 2003; Szmrécseyi 2005); ambiguity avoidance (MacDonald 1994, 1999; Clark & Fox Tree 2002; Hawkins 2004), and dependency length sensitivity (Race & MacDonald 2003; Hawkins 2004); among others.

First, in production both Bock (1986) and Szmrécseyi (2005) have looked at the effect that prior use of similar structures has on syntactic choice, i.e., priming or persistence. Bock’s work involved experiments in which spontaneous descriptions of pictures were elicited after a prime, with a finding that the form of the prime increased likelihood of a participant spontaneously producing a like form. Szmrécseyi’s corpus work corroborates this finding and also offers evidence for several parameters that modulate the effect, including textual distance of the prime, turn-taking, and how directly the priming structure relates to the evoked structure.

Bock and Warren (1985) provide empirical evidence in support of earlier hypotheses that the grammatical roles of potential referents have differential effects on production, with nominal phrases conforming to a hierarchy of relative availability and corresponding ease of production. Specifically with regard to that-mentioning, Ferreira and Dell (2000) conducted a series of experiments to compare evidence for availability and ambiguity-avoidance strategies. Their conclusion is a Principle of Immediate Mention, suggesting that it is the availability of following material that better accounts for complementizer optionality.

MacDonald (1994) used timed-reading trials to look at different types of probabilistic constraints affecting syntactic ambiguity resolution and in MacDonald 1999 further proposes that syntactic probability distributions—created online by speakers due to processing pressures but then equally available to and understood by listeners—are a thread that can unite theories of acquisition, production, and comprehension. Race and MacDonald (2003) explore additional evidence for the earlier suggestion (MacDonald 1999) that speakers and listeners draw on like probability distributions and go on to further explore dependency length, as well. Like MacDonald (1994), Hawkins (2004) also looks at both ambiguity avoidance and dependency length, proposing a principle of Domain Minimization. On the Hawkins account, inserting optional that facilitates production by reducing a dependency length in that it clearly marks CC onset. At the same time, Domain Minimization predicts that any additional words between matrix verb and CC onset increase production difficulty leading to greater probability of that-mentioning.

Returning to the influence of information density on production, nothing in the statement of UID suggests that these several other constraints and pressures on production are not also simultaneously in play. The approach taken in both the prior corpus work on UID in that-mentioning (Jaeger 2010) and the present work is to use multivariate regression to tease out these individual effects. Several predictive factors are included in the resulting statistical models to specifically control for the
simultaneous influence of alternative accounts—persistence, availability, ambiguity-avoidance, and dependency length, as well as other previously observed factors, such as demographics—enabling a view of the effect that information density has on syntactic reduction over and above these other influences.

The remainder of the present work proceeds as follows: §2 documents an initial experiment of CC that-mentioning; §3 through §5 provide evidence of the robustness of the initial results via a series of further experiments varying the rating task employed, alternation tested (RC vs. CC), and participant population engaged; followed by general discussion and conclusions in §6 and §7.

2. Testing judgments of UID

As detailed in the prior corpus works (Jaeger 2006, 2010), models fit to production data indicated that increasing information density—here meaning lower probability of complement clause embedding as conditioned by embedding (matrix) verb lemma—significantly correlated with increased that-mentioning as predicted by the UID hypothesis. In these models, the fixed effect accounting for information density made the most significant contribution to model fit while simultaneously controlling for a wide variety of other factors shown by prior research to have an effect on that-optionality. Such alternative accounts of constraints on production include availability/accessibility (Bock & Warren 1985; Bock 1986; Ferreira & Dell 2000; Race & MacDonald 2003; Szmrecsanyi 2005); ambiguity avoidance (MacDonald 1994, 1999; Clark & Fox Tree 2002; Hawkins 2004), and dependency length sensitivity (Race & MacDonald 2003; Hawkins 2004).

The experimental procedure for the present study involves soliciting participant judgments of the comparative likelihood of that inclusion or omission in various contexts. These judgments are then compared to the predictions of the corpus models.

The principal hypothesis is that mean participant ratings for that-mentioning should strongly correlate with corpus model fitted values for the same stimuli. It is further predicted that in a separate model fitted to participant ratings, information density should, as in the corpus model, make a significant contribution to model fit—while once again controlling for the simultaneous effects of other accounts. Such a result would suggest that UID plays a substantial role in judgment, just as the corpus models indicate that it does in production.

2.1. Test design

Participants are presented two alternative versions of a single transcribed utterance bearing a complement clause. In the initial study the choices are identical other than with regard to the presence or absence of complementizer that—one version with, the other without. A short written excerpt of preceding discourse
provides immediate context. The length of this excerpt varied between one and two immediately preceding sentences. The judgment task itself involves a 100-point rating paradigm suggested by the Bresnan 2007 experiments with the English dative alternation. The preceding context is noted as having been transcribed from actual recorded conversation, then participants are asked to consider how likely each of the alternatives is to have been the actual continuation, rating the choices by distributing 100 points between them.

2.2. Database

The primary goal of the study is to examine the degree to which participant judgments match predictions reported in earlier works on that-mentioning. As such, the present work was fundamentally facilitated by T.F. Jaeger graciously providing the entirety of his annotated CC database, developed through his work with probabilistic syntactic reduction (Jaeger 2006) and further employed in his more-specific look at UID (Jaeger 2010). The database contains 6,716 examples of utterances bearing a CC, programmatically extracted from the syntactically annotated Penn Treebank (release 3, Marcus et al. 1999) subset of the Switchboard corpus of spontaneous American English speech (Godfrey et al. 1992). Jaeger further programmatically annotated each token for 17 factors expected to be potentially predictive of optional that-mentioning based on earlier works on this and other syntactic reduction phenomena. (See Jaeger 2010, pps. 28-29 and 54-56, for complete descriptions of the procedures used to extract, annotate, further validate, and exclude tokens in preparation of the CC database.)

2.3. Stimuli

Test stimuli were selected from the existing database as follows. Tokens were drawn at random from the database then examined against a set of test exclusions (below). Sampling continued until a set of 30 tokens had been found to pass all exclusion tests. (The resulting stimuli are detailed in Appendix A.)

2.3.1. Exclusions

The experiment involves a written judgment task of material presented with limited context, whereas the original data are drawn from spoken, fully contextualized dialogue. Certainly, some differences may exist inherently between the factors influencing such a judgment of written material and those involved in choices during spontaneous production. To the extent possible, however, it is at least desirable to minimize overt differences, in particular avoiding tokens clearly exhibiting elements that would be unavailable or overly marked in a written task. These included exclusion of any of the following:

- Disfluencies around the complementizer site, in fact anywhere between matrix verb and CC subject
- Substantial, marked ungrammaticality in other parts of the sentence
• Repetitions or restarts at the complementizer site, for example as in (4).

(4) *I agree that, that some innocent person might be victimized by a false test.*

• Cut-off utterances where the speaker does not complete the sentence following the complementizer site
• Gender-specific language—as most tokens reveal nothing in written form about the gender of either participant in the original conversation, any stimuli that might in fact exhibit any overt markers of original speaker gender were excluded
• Proximal priming: More than one CC—and thus more than one complementizer site—in the same utterance (Given the well-documented effects of structural persistence on syntactic choice—Bock 1986; Szmrecsanyi 2005)
• Distal priming: Prior CC by either speaker within the preceding context.
• β-priming (Szmrecsanyi 2005): No related syntactic choices—meaning optional-*that* phenomena other than CCs, specifically nominal-relative clauses (RC)—by either speaker within the prior two sentences

2.3.2. Editing

Beyond the abovementioned exclusions, the resulting tokens were further edited to remove any remaining filled-pause disfluencies—*uh, um, er,* and so on—as tokens had been entirely excluded only if such elements were found immediately surrounding the complementizer site. By contrast, repetitions, restarts, and corrections using otherwise fluent words were retained. As with filled-pauses, tokens had been excluded if these issues surrounded the actual complementizer site, but they were considered acceptable within the test stimuli at other points in the utterance. Finally, long tokens with multiple conjoined sentences were split. The *S* containing the target CC became the utterance presented both with and without *that,* while the remainder of the conjoined material was retained as part of the context presented. For example, the full utterance in (5) was edited to create the stimulus shown in Figure 1.

(5) *Well, in my case my husband is not a carpenter, but in fact, he's in electronics, but he knew that the only way we'd ever have a new home is he would built it.*
2.4. Participants

Twenty-one California-resident adult native speakers of American English—twelve women, nine men, ages 23 to 63—were recruited directly by the researcher and offered $7 each for their participation.

2.5. Test procedure

Each participant saw a total of 33 stimuli. First, three practice tokens presented in the same order to each participant provided initial experience with the 100-point rating mechanism. These practice stimuli were also designed with a covert goal of encouraging a maximal spread of responses, as they were selected from the database to represent the midpoint and two extremes of predicted that-inclusion ratings, with corpus-model fitted values of 0.01, 0.51, and 0.92, respectively. Responses for these practice tokens were recorded but are not included in the statistical results. (Practice tokens are also detailed in Appendix A.)

Beyond these initial practice stimuli, the test software presented the remaining 30 tokens in a different random order for each participant, and within each stimulus the order of presentation of the alternatives—with and without optional that—was
also randomized on the fly. The software enforced the 100-point rating paradigm, requiring responses that sum to 100. A software issue apparently corrupted three individual stimulus responses out of the expected 21*30=630 ratings, yielding a final N=627 for the statistical results.

2.6. Baseline corpus model

The dependent variable in both prior corpus and current experimental studies is the binary presence or absence of optional that introducing CC. Table 1, adapted from Jaeger (2010:32), reviews the predictive factors (independent variables) present in the original corpus model. (Appendix E provides brief descriptions for each factor, but see also Jaeger 2010, pps. 30-36, for complete details.)

In particular, note that (again, per Jaeger 2010) the leading alternative prior accounts of that-mentioning introduced in §1.3 are represented in the model by three groups of factors.

- **Dependency length**
  - POSITION (MATRIX VERB)
  - LENGTH (MATRIX VERB-TO-CC)
  - LENGTH (CC ONSET)
  - LENGTH (CC REMAINDER)

- **Availability/accessibility**
  - CC SUBJECT—form of nominal
  - SUBJECT IDENTITY—CC subject identical to matrix subject
  - FREQUENCY (CC SUBJECT HEAD)
  - WORD FORM SIMILARITY—CC subject begins with demonstrative that
  - FREQUENCY (MATRIX VERB)

- **Ambiguity avoidance**
  - AMBIGUOUS CC ONSET

2.7. Adjustments to baseline corpus model

Before comparing the experimental ratings to corpus predictions it is necessary to adjust the corpus model slightly and recalculate its fitted values—its predictions for that-mentioning, hypothesized to correlate with experimental judgment ratings. Specifically, several factors that figure in the published CC corpus model (Jaeger 2006, 2010) represent variable elements which—while available to speakers during the recorded telephone conversations that comprise the Switchboard corpus—are not available to survey participants in the present judgment of written transcriptions. The unavailable bits include:

- Pauses—unfilled silence at any point in the utterance, most critically any such pauses surrounding the complementizer site.
• Disfluencies—filled pauses, such as *um, uh, er*, and so on. Recall that stimuli evincing these elements surrounding the complementizer site were excluded, and where present in other parts of an utterance selected as a stimulus, such filled pauses were edited out.

• Persistence—Tokens were excluded as stimuli if they exhibited any form of *that*-choice priming within the discourse context provided—whether overt priming by another CC or indirect priming by a RC construction ("β-persistence," cf. Szmrecsanyi 2005).

• Speech rate—While an approximation of participant reading rate might appear to function as something of a proxy for production speech rate, these elements are not in fact related. Ultimately, experiment participants do not have access to information about the rate of speech of the original utterance and cannot reflect such in their ratings of examples presented in written form.

• Original speaker gender.

The original corpus-based predictive regression model was reconstructed and refit with these elements removed. The comparison below of experimental ratings against corpus predictions is based on this revised model. Of the several variants discussed in Jaeger 2010, the reconstructed model employed here is based on the version that includes a random effect of matrix verb lemma to account for potential grammaticalization in the form of any per-verb bias for *that*-mentioning separate from the main study variable of conditional probability of CC in context. As with the original (Jaeger 2010:29), the model was refit here using Laplace Approximation as implemented in the R software package *lme4* (Bates et al. 2008).
Table 1, adapted from Jaeger (2010:32), reviews the factors present in the original corpus model, highlighting those factors now removed in refitting the model for the present work.

Table 1: Predictors in the original corpus model (adapted from Jaeger 2010:32). The rightmost column gives type and number of model parameters. Grayed areas indicate those factors that the present work removes for comparing corpus model predictions to judgments from the current study’s written-task experiment.

Table: Predictor Description Type

**RANDOM**
- SPEAKER: ID of current speaker, rand
- LEMMA(MATRIX VERB): Lemmatized form of matrix verb, rand

**FIXED**
- INTERCEPT
- INFORMATION(CC ONSET): -log conditional probability of CC onset in context, cont(1)
- POSITION(MATRIX VERB): CC position in the sentence, cont(3)
- LENGTH(MATRIX VERB-TO-CC): Distance of CC from matrix verb, cont(1)
- LENGTH(CC ONSET): Length of CC onset, cont(1)
- LENGTH(CC REMAINDER): Length of remainder of CC, cont(1)
- SPEECH RATE: Log and squared log speech rate, cont(2)
- PAUSE: Pause immediately preceding CC, cat(1)
- DISFLUENCY: Normalized disfluency rate at CC onset, cont(1)
- CC SUBJECT: Type of CC subject, cat(3)
- SUBJECT IDENTITY: Matrix and CC subject are identical, cat(1)
- FREQUENCY(CC SUBJECT HEAD): Log frequency CC subject head lemma, cont(1)
- WORD FORM SIMILARITY: Potential for double that sequence, cat(1)
- FREQUENCY(MATRIX VERB): Log frequency of verb lemma, cont(1)
- AMBIGUOUS CC ONSET: CC onset ambiguous without *that*, cat(1)
- MATRIX SUBJECT: Type of matrix subject, cat(3)
- SYNT. PERSISTENCE: Prime (if any) w/ or w/o *that*, cat(2)
- MALE SPEAKER: Speaker is male, cat(1)
2.8. Results

Multiple statistical procedures will be used to evaluate the degree to which participant judgments align with corpus model predictions.

2.8.1. Ratings correlation

At the highest level, we examine to what extent the ranked mean experimental that-mentioning ratings for the 30 stimuli correlate with the corresponding ranked that-inclusion predictions (fitted values) of the corpus-derived regression model. Figure 2 illustrates this correlation. (Appendix B details ratings for each token.)

![Figure 2: Ranked mean ratings for each token (y-axis) plotted against ranked corpus model fitted values (x-axis). The 45°-angle line would represent perfect alignment between predictions and results.](image)

Visually, the non-parametric (Lowess) smoothed line among the ranked mean ratings aligns extremely well, if not perfectly, with the ranked corpus predictions. Statistically, the rank correlation (Spearman’s \( \rho \)) is more appropriate than a Pearson
product moment correlation in that neither predictions (corpus model fitted values) nor mean ratings are normally distributed, as is assumed for a Pearson’s test. Confirming the visual impression, participants do in fact successfully predict ($\rho=0.800$, $p<0.0001$) speaker *that*-production in introducing complement clauses, against a model where the greatest contribution to fit comes from the UID parameter, representing Shannon information content associated with CC onset in context. Figure 3 illustrates the raw (non-ranked) mean ratings.

![Figure 3: Mean (non-ranked) ratings (y-axis) against corpus model fitted values (x-axis). The 45°-angle line represents perfect alignment between predictions and results.](image)

Intuitively, one might initially view the cluster of points to the right of 0.4 on the x-axis as outliers in terms of participant ratings, but it is important to observe that while these points are noticeably offset to the right on Figure 3, they are not offset vertically. As the x-axis represents the corpus prediction for each stimulus, the right offset here merely illustrates that in the random selection of tokens from the database, there was a gap in the sample. This is less surprising on observing the comparative sparseness of corpus fitted values above 0.2. (See Figure 10.)
With this noted, we do see that overall, experimental ratings appear to cluster higher and with a somewhat narrower spread than predictions. This is confirmed in a comparison of the grand means—0.44 for the experimental ratings against 0.17 for the corpus model fitted values—and standard deviations among points—σ=0.12 for ratings against σ=0.18 for corpus predictions. The nature of the 100-point judgment paradigm or written task may tend to encourage less divergence. Or even while highly correlated overall, the somewhat compressed range may reflect a partial difference between the cognitive processes of judgment and production—while still showing the overall correlation expected by the study.

2.8.2. Modeling judgment

We may achieve additional confidence in the alignment between judgment and production by investigating the factors underlying the experimental ratings. If participants not only reliably predict speakers’ ultimate choices in CC that-mentioning but in fact turn out to do so based on similar probabilistic factors, this would provide evidence of an even stronger alignment between judgment and production.

To investigate, a multilevel generalized linear mixed model (Baayen 2008) was fit to the experimental results, with the 100-point ratings as the dependent (predicted) variable. An obvious difference between this judgment model and the corpus-based production model is that the latter predicts a binary choice—with or without that—modeled according to a binomial distribution, whereas the 100-point judgment ratings produce gradient results, modeled here using a Normal, or Gaussian, distribution. Otherwise, the fixed-effect predictors (independent variables) match those used in the corpus model that was refit for the present study (Table 1). The random effect of corpus speaker is removed and replaced by a similar random effect based on participant ID to control for individual that-mentioning bias among experiment participants. To this, a handful of additional fixed-effects controls—noted in Table 2—are added to account for variables in the experimental context.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RANDOM</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARTICIPANT</td>
<td>Individual experiment participant</td>
<td>rand</td>
</tr>
<tr>
<td><strong>FIXED</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GENDER(MALE)</td>
<td>Judgment from male participant</td>
<td>cat(1)</td>
</tr>
<tr>
<td>AGE</td>
<td>Participant age</td>
<td>cont(1)</td>
</tr>
<tr>
<td>ORDER</td>
<td>Current token's position in random sequence (1-30)</td>
<td>cont(1)</td>
</tr>
<tr>
<td>THAT FIRST</td>
<td>Alternative w/ that presented 1st in current token</td>
<td>cat(1)</td>
</tr>
</tbody>
</table>

Table 2: Additional predictors reflecting experimental factors, added to the regression model fit to experiment ratings.

A first optimization of the resulting model is made by removing the random effect for matrix verb lemma. While this factor significantly improved model fit in
the corpus production model, it yields no improvement at all in fit—as measured by restricted maximum likelihood (REML)—against the judgment data, perhaps unsurprising given the much smaller sample of verbs involved among the 30 experimental stimuli. Removing this variable improves the model’s Akaike information criterion (AIC) evaluation, since the variable does not improve REML, and between models with like deviance, AIC favors a model with fewer predictors (fewer degrees of freedom).

2.8.2.—1. Examining the new experimental factors
   2.8.2.—1.1. Gender

   In the original corpus production model, speaker gender was included but did not achieve significance, with women found to be marginally more likely to mention *that* than men. In an initial univariate analysis of the judgment data, the trend seems to be reversed, with men significantly \((p<0.0001)\) more likely to predict *that*.

2.8.2.—1.2. Age

   An initial look appears to indicate that increasing participant age correlates with decreased *that*-mentioning, though not significantly predictive on a univariate basis \((p>0.5)\). However, visualizing via a non-parametric smoother in Figure 4 we see what appears to be a more complex effect, where up to about age 40, increasing age correlates with increased *that*-mentioning, then the trend sharply reverses above age 40, with even perhaps a turnaround again above age 50. The straight dotted line illustrates the simple regression, but the full curve is suggestive of a quadratic ('U'-shaped) or even cubic ('S'-shaped) effect. Indeed, including squared and cubed terms improves fit (explains more variance), with the \(<age^2>\) factor at least marginally significant \((p=0.08)\), while the linear effect of age remains not so. This noted, no explanation for such variation by age is immediately clear.

![Figure 4: Participant age as a predictor of *that*-inclusion judgment. The dotted line represents a univariate linear regression.](image)

---

15
2.8.2.—1.3. Order

Position in the randomized sequence is marginally significant on univariate analysis ($p=0.08$), with *that*-mentioning decreasing slightly on average as participants proceed through the 30 stimuli.

2.8.2.—1.4. *that*'-first effect

Interestingly, the order in which the alternatives were presented within a given stimulus is strongly significant ($p<0.0001$) on univariate analysis, with *that*-inclusion judged more likely when the first alternative included *that*—seemingly the power of initial suggestion.

2.8.2.—2. Collinearity

Using a $\kappa$ condition calculation, collinearity among factors does not appear to be a concern—neither among the new factors related to the experimental context ($\kappa =1.8$) nor for the full model ($\kappa =4.3$). (A $\kappa$ condition value of 0 to 6 is considered within reasonable limits of collinearity; Baayen 2007:182).

2.8.2.—3. Comparing judgment and production effects

Returning our attention to the primary research questions, the demonstrated ability for participants to accurately predict speaker choices—as seen above in the strong correlation between net judgment ratings and corpus model fitted values—already suggests that the factors underlying production, including UID, are also involved in judgment. With a full regression model now fitted to experimental judgments, however, we can look deeper.

2.8.2.—3.1. Factors affecting judgment

Table 3 summarizes both the judgment model and the revised corpus model refit as described above to remove elements not accessible to participants in a written judgment trial. Estimated $\beta$ coefficients and $p$ levels of significance are given for all factors.
Table 3: Judgment and production models, with estimated $\beta$ coefficients and $p$ significance levels. Judgment model includes the additional factors accounting for experimental context. Production model is a reduced and refit version of the models described in Jaeger 2010.

We can immediately observe that some of the additional variables encoded to account for the experimental environment and context—the dotted-line enclosed box in Table 3—are significantly influential in the test results. While controlling for all other factors, the participant gender effect—MALE PARTICIPANT indicating a higher preference for *that*-inclusion—remains significant, as does the influence on *that*-mentioning of having the stimulus alternative bearing *that* presented first in the pair of sentences. Finding *that*-inclusion in the first alternative (THAT FIRST) significantly generates higher ratings for the likelihood of *that* having been included in the actual discourse.

On the production (corpus) side, with several factors removed from the full corpus model, estimated coefficients and $p$ values are obviously somewhat changed, but of the remaining factors, all that were significant in the full model remain
significant here, and conversely, no factors that had been less than significant in the full model rise to significance here.

2.8.2.–3.2. Discordant effects

Three factors on Table 3 undergo a change in polarity between the models: SUBJECT IDENTITY, LENGTH(MATRIX VERB-TO-CC), and MATRIX SUBJECT=OTHER PRO.

Of these, the first two may simply reflect clear differences between natural speech production and transcribed speech evaluation processes. In the corpus, SUBJECT IDENTITY—which indicates string-identical (and assumed co-referential) matrix and CC subject—significantly reduces that-mentioning, taken as evidence in support of an availability-based account of production. Already primed by its co-referent, the CC subject should be highly accessible, enabling quicker production of CC onset and less need for delay via that-insertion. Since the present judgments are made only after reading and processing the full multi-sentential stimuli, such availability pressures should be greatly reduced, which may account for this factor’s lack of significance in our judgment model, as well as its reversed polarity.

By contrast, where availability accounts (Bock & Warren 1985; Ferreira & Dell 2000; inter alia) expect subject co-referentiality to ease production pressure and reduce that-mentioning, dependency length accounts (Hawkins 2004) predict that words intervening between matrix verb and CC onset—noted in our model via LENGTH(MATRIX VERB-TO-CC)—should increase production difficulty and thus that-mentioning. In our offline judgment task without real-time processing demands, however, this effect should be substantially diminished, which may account for its opposite polarity here.

The reversed polarity of MATRIX SUBJECT=OTHER PRO, on the other hand, is not quite as clearly motivated, though it may be related to markedness, via both discourse pragmatics and lower lexical frequency. The MATRIX SUBJECT=OTHER PRO factor was encoded in the Jaeger 2010 database as all personal pronouns other than I and you (Jaeger 2010:35). The Switchboard corpus, however, is comprised entirely of live one-on-one telephone conversations. Pragmatically, the bulk of speech can be expected to be in the first- and second-person singular. First-person plural we and any third-person pronominal subjects would seem to be substantially marked for such discourse. This is further borne out by a quick check of each form’s frequency\(^1\) in subject position in Switchboard; I (103K tokens) and you (59K) are indeed most frequent, with only neutral it (50K) coming anywhere close, and he, she, they, and we

---

\(^1\) It may well be that pragmatic markedness and low-frequency markedness are simply two sides of the same coin, with one in fact leading to the other, or vice versa. Indeed, it has been argued elsewhere that most, if not all, types of linguistic markedness can be attributed almost entirely to frequency (Haspelmath 2006). For present purposes, however, this question need not be answered here. It is sufficient to note that regardless of cause, subject pronouns other than I and you are marked in the production data context in a way that they are not for the present work’s judgment participants.
well behind. The increased processing required to identify a referent other than oneself or one’s immediate interlocutor may be expected to increase either processing load or competition for cognitive resources or both, leading to delayed CC onset via that-mentioning. As experiment participants are inherently third-party observers here, such markedness may be substantially diminished, which would account for the experiment data’s greatly reduced bias towards that-mentioning in non-I or you subject contexts.

2.8.2.—3.4. Effect of Uniform Information Density

We may now return attention specifically to information-theoretic concerns and the predictions of Uniform Information Density. As noted in the introduction, in the prior work on that-mentioning and UID (Jaeger 2010) the information content \( I \) represented by a CC onset, or \( I(CC|context) \approx -\log p(CC|matrix \ verb \ lemma) \), made the greatest contribution to model fit. While simply finding that experimental ratings match corpus model predictions suggests that UID is in play in judgment, it is worth making a comparison specifically of the information density effects in the two models.

The polarities match, indicating that, as predicted, increased information content—greater “surprisal”—predicts more likely that-mentioning, explained under UID as an attempt to smooth out this information peak by spreading the information across more linear words.

Finally—and again, while not strictly required to show that UID plays at least some role—it is worth noting that among all factors, the information density effect in fact makes the greatest single contribution to model fit in judgment, just as it does in the corpus production model. Specifically, the UID parameter (INFORMATION DENSITY) accounts for a delta of -44 in AIC model-fit evaluation, greater than any other single factor.

3. Seeking breadth: Additional trials

Where §2 looked in depth at the alignment of production and judgment processes with regard to UID, the following sections seek breadth to complement this depth. The robustness of alignment is explored by repeating the initial experiment while varying the test design along three different dimensions: the rating task employed, the alternation tested (RC vs. CC), and the participant population engaged.

3.1. Alternative judgment task: speeded binary forced-choice

While the initial experiment using a 100-point rating paradigm (after Bresnan 2007) provides a strong complement to corpus work on UID, post-experiment

---

2 As measured by extractions via TGrep2 software (Rohde 2005).
interviews led to a suspicion that participants were to some extent “over-thinking” their ratings. I considered that a speeded task might yet induce a greater spread of responses.

3.2. Test design

The same 30 test stimuli from Experiment 1 (100-point rating) are used, with new participants seeing the same three practice questions in the same order before proceeding to a randomized presentation of the remaining stimuli. Rather than allocating 100 points between the choices, participants were asked to select the alternative they considered more likely to have been the actual continuation of the discourse. An on-screen countdown timer indicates how much time remains to make a selection. Figure 6 shows an example task.

![Figure 6: Example stimulus with preceding context and selection mechanism. In this screen capture, the countdown timer indicates eight seconds remaining.](image)

When the timer expires, the test software moves on to the next randomly ordered stimulus without accepting a response for the current item. Timer length was derived separately per item as a linear function of the number of words in the stimulus. The parameters of this function were tuned via a small initial pilot with the goal of being sufficiently short as to distinctly hurry the responses but not so
short that any substantial portion of the tasks would be left with a null response. An average loss of 10%—three items per participant—was the initial target. The resulting timer lengths ranged from six to 16 seconds.

3.3. Participants

Twenty-nine California-resident adult native speakers of American English—twenty women, nine men, ages 23 to 62—who had not participated in the first experiment were recruited directly by the researcher and offered $7 each for their participation. Thirty stimuli each yields 29*30=870 trials.

3.4. Spoilage

Timer expirations removed 52 trials, or 6% of the 870 potential judgments, for a mean loss rate of 1.8 items per participant, or viewed alternatively, 1.7 participants per item. No participant had more than five expirations, for a maximum loss rate of 16.7% of 30 items. Maximum expirations for a single stimulus were seven, or 24% of 29 participants.

3.5. Results

Analysis proceeds as for the previous 100-point trial. (Appendix B details ratings for each token for all experiments.)
Figure 7: Ranked mean forced-choice (y-axis) against ranked corpus model fitted values (x-axis).

Participants once again successfully predict speaker that-production, with strong correlation (Spearman’s rho, $\rho=0.716$, $p<0.0001$) to a corpus model in which information density makes the greatest contribution. Previously, however, it was the raw (not rank-ordered) means where the 100-point task seemed to yield a comparatively narrow range of ratings. Figure 8 shows mean forced-choice ratings side-by-side with the earlier 100-point ratings (from Figure 4, §2.8.1, repeated here as the left-hand panel on Figure 8).
The speeded forced-choice here yields a wider spread of mean ratings than the prior 100-point experiment. Mean rating (mean of means) is almost unchanged—0.42 for forced-choice against 0.44 for 100-point—but standard deviation climbs from $\sigma=0.12$ for 100-point (left panel) to $\sigma=0.17$ for forced-choice (right panel), nearly reaching the spread ($\sigma=0.18$) predicted by the corpus model (fitted values).

3.5.1. Modeling

Beyond the replication of the surface correlation, a brief replication on these forced-choice data of the model analysis in §2 for the 100-point data may provide still further evidence for robustness of the production-judgment alignment previously observed.

Once again a mixed-effects linear regression was fit to the judgments, now using a binary include-vs.-omit-\textit{that} choice as the dependent variable. As such, this is a logistic regression with binomial distribution rather than the Gaussian (Normal) model fit to 100-point judgments. Modeling otherwise proceeds as before, with the same additional predictors to control for variables in the experimental context, including age and gender of participants, order of presentation within the random sequence for each token, and whether the \textit{that}-bearing alternative appears first. Collinearity remains low, with $\kappa$ condition of $\kappa=4.6$.

In terms of effects obtaining from the experimental context, \textit{that}-inclusion in the first alternative (\textsc{that first}) continues to significantly correlate with higher ratings for the likelihood of \textit{that}–mentioning ($p<0.0001$), as it did with the 100-point task.
Curiously, the male participant bias for *that*-mentioning is quite insignificant here \((p>0.8)\) despite having drawn participants from a similar population, while the intriguingly cubic ('S'-shaped) participant age effect, \(<\text{age}^3>\), is actually more pronounced here, rising to significance even under multivariate control \((p=0.050)\).

### 3.5.3. Discordant effects

Similar to what we observed for the 100-point task, \textsc{subject identity} and \textsc{matrix subject}=\textsc{other pro}, remain prominent outliers, joined here as well by \textsc{cc subject}=\textsc{it}. The last of these seems likely to have a motivation similar to that previously suggested for \textsc{matrix subject}=\textsc{other pro}, where experiment participants—being post-hoc, third-party observers—may find matrix and embedded subject nominals other than \textit{i} and \textit{you} less marked than they appear to have been for the original telephone interlocutors modeled by the corpus-based regression. (See §2.8.2 for the prior discussion.)

### 3.5.4. Uniform Information Density

Finally, examination of individual effects confirms that as with the 100-point task, the \textsc{uid} parameter (\textsc{information density}) continues to make the single greatest contribution to model fit, as measured by AIC.

### 3.6. Discussion

Table 4 offers an interim summary of results. We find that the forced-choice task induced a broader spread of responses (greater standard deviation) than the 100-point paradigm, more in line with the corpus production data. At the same time, multivariate regression modeling for this second experiment reconfirms the deeper alignment of production and judgment for *that*-mentioning, including the contribution of information density to the syntactic reduction.

<table>
<thead>
<tr>
<th></th>
<th>100-points</th>
<th>forced-choice</th>
<th>corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean rating (μ)</td>
<td>0.44</td>
<td>0.42</td>
<td>0.17</td>
</tr>
<tr>
<td>Standard deviation (σ)</td>
<td>0.12</td>
<td>0.17</td>
<td>0.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>100-points</th>
<th>forced-choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation of mean judgments against corpus predictions (ρ)</td>
<td>0.80***</td>
<td>0.72***</td>
</tr>
<tr>
<td>Correlation of judgment model factors against corpus model (ρ)</td>
<td>0.51*</td>
<td>0.54*</td>
</tr>
</tbody>
</table>

\textbf{Table 4:} Summary comparison of CC *that*-mentioning judgment trial results.
4. Seeking breadth II: Relativizer omission

Intuitively, *that*-mentioning behaves similarly for relative and complement clauses, and Jaeger 2006 shows that these syntactic-reduction alternations are similarly predictable from context. Where the probability estimate for CC onset in §2 and §3 was predicated on embedding verb lemma, for RC onset Jaeger similarly estimates contextual probability based on relativized head noun lemma. One way in which RC and CC *that*-reduction models differ, however, is in the shape of the probability distributions they predict, as illustrated in Figure 10.

![Figure 10: Corpus-model prediction (model fitted value) of *that*-inclusion likelihood (y-axis) for 100-token random samples from CC and RC databases.](image)

While corpus-model predictions of *that*-mentioning for CCs cluster towards zero (*that*-omission), with only a small number of outliers anywhere above 0.5 likelihood of including the complementizer, corpus-model-predicted likelihoods of *that*-mentioning for RCs are more broadly spread, forming a quite even gradient from near zero (categorical *that*-omission) to nearly 1.0 (categorical *that*-inclusion). With less clustering, the RC model may be seen as a more difficult target—or setting a higher bar—for a judgment trial of *that* likelihood to attempt to match. If participant ratings of RC relativizer *that*-mentioning likelihood match corpus-model predictions in a manner similar to that found for CC complementizers in §2 and §3 it would provide still further evidence of a robust linkage between production and judgment processes for information theoretic syntactic reduction.

Both the 100-point rating (§2) and timed forced-choice (§3) tasks were repeated for RC data.
4.1. Database

Stimuli for these experiments were developed from a database from prior work on syntactic reduction by T.F. Jaeger (2006). Like his CC database, the RC tokens were programmatically extracted from the syntactically annotated Penn Treebank (Marcus et al. 1999) subset of the Switchboard corpus (Godfrey et al. 1992). The RC database includes a total of 3,048 tokens.

4.2. Stimuli

Thirty test stimuli were developed using the same exclusion and editing procedures as described in §2.3 for the CC stimuli. The only substantial difference is that rather than drawing each token at random from the entire database, broad coverage of the wide gradient of predicted that likelihoods was ensured by first dividing the database into five probability bins—0.0-0.2, 0.2-0.4, etc.—then drawing tokens until, after exclusions, there were six stimuli from each of the five bins.

4.3. Participants

For the 100-point trial, 11 female and 14 male Stanford undergraduates, ages 18 to 23, were recruited by a research assistant, with one additional male non-student also participating. For the forced-choice trial, 17 women and 11 men, ages 19 to 60, were recruited by the researcher from the general population. All participants were paid $7.

4.4. Adjustments to baseline corpus model

As described in §2.6 for the CC experiments, the corpus that-prediction model for RCs was similarly refit after excluding conditions not observable by participants in a written judgment task, including pauses, disfluencies, persistence, speech rate, and original speaker gender.

4.5. Forced-choice timers and spoilage

As in §3.3 for CCs, the forced-choice portion of the RC experiments involved countdown timers set so as to hurry the task. Timer lengths were once again set as a linear function of stimuli length, with resulting times ranging from five to 24 seconds. Timer expirations removed 63 of 840 interactions. However, a single stimulus out of 30 accounted for 12 of these expirations, with no other single stimulus losing more than five judgments. It would seem the timer was inappropriately short for this one item. After removing this outlier, there were 51 expirations out of 812 interactions (6.3%). By participant, this is an average of 1.96 out of the remaining 29 stimuli. The mode is zero, with maximum (one participant) six. By stimulus, after removing the outlier, there were an average of 1.76 participants eliminated, with minimum zero and maximum five.
4.5. Results

Figure 11 illustrates both rank and non-ranked correlations between participant ratings and corpus-model predictions for both 100-point and forced-choice tasks.

**Figure 11**: Upper panels show rank-order correlations for judgment ratings (y-axes) against corpus-model predictions (x-axes) for relativizer *that*—mentioning. Lower panels show direct (not rank-ordered) comparisons.

Methodologically, we see again that as in the CC experiments, the forced-choice paradigm induces a greater spread of mean responses (lower-right panel vs. lower-
The resulting separation of responses in turn also yields a tighter rank-order correlation between participant ratings and corpus-model predictions (upper-right panel vs. upper-left).

Method aside, both RC tasks show highly significant correlations ($\rho=0.64$, $p<0.001$ for 100-point distribution; $\rho=0.71$, $p<0.0001$ for speeded forced-choice), offering still further evidence of the robustness of relationship between production and judgment processes.\(^3\)

### 5. Seeking breadth III: Population diversity

Having shown that this correlation is robust across variations both in terms of task (100-point vs. speeded forced-choice) and alternation phenomena (CC onset vs. RC onset), a final measure of confidence in the results is sought by varying the population queried. Here, crowdsourcing—ratings by a large number of anonymous participants via the Internet rather than known participants in traditional lab experimentation—offers an opportunity to check our results with a more highly diverse population.

Methodologically—if orthogonal to the primary goals of the present work—this does also afford an opportunity to examine whether traditional lab experiments can be successfully replicated using an anonymous Internet population.

To this end, all four of the previously discussed experimental variations—RC vs. CC; 100-point vs. forced-choice—were replicated via crowdsourcing.

#### 5.1. Test procedure

The crowdsourced experiments were conducted using the Amazon Mechanical Turk (AMT) online Human Intelligence Task (HIT) mechanism. Stimuli were advertised and presented on AMT as individual tasks, as illustrated in Figure 12.

\(^3\) Unlike the CC experiments, a fully annotated RC database was not available here for modeling and deep comparison of underlying factors. Also, within the experimental context, only age and gender were recorded, not token order after randomization or which alternative included that for each token. Fitting mixed-effects models to the ratings, with participant as a random effect to control for per-individual variation/bias, indicates that neither age nor gender had a significant effect on results in these trials.
Unlike typical controlled lab experiments, one aspect of the way the tokens were presented to AMT is that different participants would not necessarily see the same stimuli nor respond to the same number of stimuli, resulting in a highly unbalanced data set. It might be anticipated that among other effects, participants might develop different biases for that-mentioning through the course of their exposure to the trial. Statistical controls in the processing of results compensate for this imbalance.

Specifically, for each of the two sets of stimuli, CC and RC, two batches were created, each containing all 30 of the tokens but placed in two different orders. Within this, the order of the that-inclusion alternatives was also varied within each batch. Each batch was then submitted to AMT for 20 different participant judgments per token, yielding a total of 1,200 individual judgments (30 x 2 x 20). For each of the four trials, its two batches were submitted on different days and at different times of day to increase the chance of being responded to by as close to a disjoint set of participants as possible. Controls within the AMT mechanism ensure that no single token was seen more than once by the same participant within a given batch. It was, however, possible for a first-batch participant to return and respond to the same token(s) in the second batch. For example, in the 100-point task,
complementizer experiment, the 1,200 judgments were contributed by a total of 104 participants. Figure 13 presents the distribution of number of tokens per participant for this experiment, summed across the two batch submittals.

![Histogram of tokens responded to per participant](image)

**Figure 13:** Histogram of tokens responded to per participant

As Figure 13 immediately reveals, four participants contributed more than 30 judgments by apparently returning to see some number of tokens for a second time with the second batch submission. Further programmatic filtering uncovers a total of seven participants who contributed a second judgment for one or more tokens in this manner. All such second, or return, judgments were excluded.

### 5.2. A Priori Filtering

As in any experimental procedure, some care must be taken to ensure the integrity of responses. Uncooperative participants might answer at random in order to rush through the process. Subjects may engage in “satisficing” (Simon 1956), providing only responses that they think will ensure they get paid, rather than truthful answers to demographic questions or fully thoughtful reactions to the stimuli. Given the anonymous nature of crowdsourced studies, however, these concerns are even greater than for face-to-face lab trials. Without bias to the primary judgment response, several further a priori controls were implemented to mitigate the increased variance we might otherwise expect due to such non-cooperation factors.
5.2.1. 100-point sum

With the two 100-point trials (RC and CC), the first and most obvious control was simply to reject any items where the points distributed between the alternatives did not sum to 100. This removed an average of 2% of the judgments. (The lab version of the test software embedded code to force a valid response, but for these anonymous crowdsourced experiments it was deemed better to filter inattentive responses than to try to force attentiveness.)

5.2.2. Inconsistent demographics

Responses with no answer for gender or age and in one case a single response with a nonsensical age (“2!”) were removed. Across the four experiments the remaining participant ages ranged from 17 to 58. Further, in some cases participants contributed judgments for more than one stimulus but provided inconsistent responses to either age or gender or both across their multiple tokens, from which we may infer that their responses were random, or at best inattentive. (One participant provided no fewer than 16 different values for age!) All judgments by these participants were removed.

5.2.3. Language and location

The present studies target adult native speakers of American English, but unlike a lab study, crowdsourced participants cannot be as easily pre-selected to match demographic considerations. The larger concern once again, though, is satisficing. Given the anonymity of the process, if the survey were to indicate that participants must match certain parameters, respondents might be all too likely to falsely claim membership in the target group. The compensating technique employed in the present study was to not pre-screen participants, paying all comers, then asking demographic questions without indicating that there was any preferred response. Responses from those indicating that English was not their first language were then removed. As for targeting American English—which would require knowing where participants were raised—asking where respondents currently live was used as a proxy, removing an average of 268 non-U.S. judgments per trial. As with age and gender above, all judgments were also removed from participants who gave inconsistent responses to these questions across multiple tokens.

5.2.4. Response time

One further difference from the lab test procedure is that no countdown timer was used on AMT, even on the forced-choice trial. Whereas in the lab it was speculated that hurrying participants might lead to better alignment between experimental results and corpus-model predictions, crowdsourcing workers have an inherent economic incentive to move as quickly as possible through each task, since they are paid per individual judgment. With the crowdsourced studies, in fact,

---

4 Herb Clark (personal communication) quite rightly points out, however, that in asking if English was a participant’s first language, the nature of the yes/no question inherently telegraphs the likely desired response, almost certainly leading to some degree of satisficing. Future versions of this task should instead pose the more open-ended, “What was your first language?”
response time—automatically collected via the AMT mechanism—emerges as another possible indicator of non-cooperation. Overly quick responses likely indicate “click-through” behavior, while excessively long times suggest either a distracted participant or one who is giving the decision too much thought. A common approach in psychology is to remove at least the high end of response times as an indicator of disproportionate difficulty in processing (cf. Winawer et al. 2007) or, in many psycholinguistic judgment studies, anything lying beyond 2.5 standard deviations from the mean in either direction, both the high and low 0.6% assuming normally distributed data (cf. Blackwell et al. 1996; Hofmeister & Sag 2010; inter alia). This avoids outliers having an overstated effect in regression measures (Ratcliff 1993).

As Figure 14—an example from the 100-point CC crowdsourced experiment—illustrates, however, the response times in the present studies are clearly not normally distributed. A large tail of very long times—up to 110 seconds, or nearly two minutes—inflates the standard deviation (sigma). At the low end, though, with such a large standard deviation relative to the mean—σ=16.7 seconds in this example, against μ=30.4 seconds—the entire lower tail lies within just 1.69σ. A 2.5σ cutoff would thus have no effect at all, and yet considering that four demographic questions must be answered before the stimulus is even considered, any total response time less than 10 seconds seems entirely implausible.

![Figure 14: Distribution of response times (CC, 100-points study). Times reported include responding to the demographic questions in addition to the judgment task.](image-url)
The hybrid approach seems appropriate here. Across the four crowdsourced studies, responses under 10 seconds were dismissed—3.0% of the data in the CC 100-points example—then at the high end, common practice was followed by pruning beyond the 2.5σ mark—corresponding with 3.4% of the long-response data for the CC 100-points example.

5.2.5. Net effect of exclusions

In the CC 100-points example, the complete set of a priori exclusions discussed above resulted in removal of 459 responses (38.3% of 1200) for the CC 100-points study, yielding a working database of 741 judgments.

5.3. Results

Figure 15 (following page) shows the rank-order mean participant ratings vs. corpus-model predictions (judgment vs. production) comparisons for all four crowdsourced experiments, with all in fact yielding significant correlations, as they did with the lab populations.

5.4. Discussion

Table 5 summarizes all of the correlations we have seen now across the set of eight experiments.

<table>
<thead>
<tr>
<th></th>
<th>Lab</th>
<th></th>
<th>Crowdsourced</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100-points</td>
<td>forced-choice</td>
<td>100-points</td>
<td>forced-choice</td>
</tr>
<tr>
<td>CC</td>
<td>0.80***</td>
<td>0.51*</td>
<td>0.60***</td>
<td>0.63***</td>
</tr>
<tr>
<td>RC</td>
<td>0.64***</td>
<td>0.71***</td>
<td>0.56**</td>
<td>0.38*</td>
</tr>
</tbody>
</table>

Table 5: Summary comparison of CC that-mentioning judgment trial results.

Crowdsourced results can generally be expected to be “noisier” (greater variance) given the unbalanced data sets involved, and non-cooperation remains a concern. It is perhaps unsurprising then that three of four correlations illustrated in Figure 15 are degraded somewhat (lower significance) than their counterparts among the lab studies, though again, each does still reach significance.

These complementary results with a generally diverse (if anonymous) population offer a final round of evidence for the robust nature of the suggested alignment of judgment and production in information-driven syntactic reduction.

Methodologically, these studies indicate that lab results can indeed be substantially replicated using an anonymous Internet population.
Figure 15: Rank-order correlations for judgment ratings (y-axes) against corpus-model predictions (x-axes) for all four crowdsourced experiments.

6. General discussion

The present study involved a set of experiments to explore the effects of information density on judgments of the likelihood of optional syntactic elements, in line with the predictions of the Uniform Information Density (UID) hypothesis. As such, this work complements existing corpus-based evidence for UID.
Consistently, experiment participants were able to accurately predict speakers’ production of optional *that*. Specifically, participants’ ratings in judgment paradigms significantly correlate with the predictions (fitted values) from models of complementizer *that* production derived from a corpus of spontaneous speech, indicating that the same factors involved in production are also involved in judgment. This alignment further suggests that sensitivity to information density is part of speaker and listener’s implicit knowledge of their language. While still further explanations for this alignment may as yet be possible, the burden of proof would seem to be on those who would propose such alternatives to account for the present results.

Further, by fitting multivariate models to these judgment and production data it was found that, while simultaneously controlling for several other accounts of influence on *that*-mentioning suggested by other prior research, information density makes the single largest contribution to model fit—to explaining the variance involved in the *that*-mentioning alternation phenomena—in judgment as it does in production. (Strictly speaking, finding that UID makes this standout contribution with this particular phenomenon was not even entirely necessary to make the case that UID is involved to *some* degree in both production and judgment and is thus part of speaker implicit knowledge.)

Next, having shown these results for a gradient 100-point rating paradigm, findings were replicated using a speeded forced-choice task. Additional breadth of support comes in seeing that judgments similarly match predictions for an information-driven model of syntactic reduction with relativizer *that*, as well. Finally, a complete replication of all lab studies using an anonymous Internet (crowdsourced) population yielded substantially similar results, providing even greater confidence in the robustness of the observed alignment of information-driven judgment and production.

Naturally, the results reported above do leave some questions unanswered. In particular it was noted that of the more than a dozen factors found significant in the corpus-derived production, only a few of these were significant in predicting experiment participants’ judgment ratings. Why might this be? A number of possibilities arise.

First, of course, is the obvious difference in task. While the present study finds evidence of a correlation between offline judgment and online production in syntactic reduction phenomena, along the lines previously shown with other alternation phenomena (Bresnan 2007; Bresnan & Ford 2010), there is certainly no suggestion here that judgment and production should behave entirely identically with regard to these alternations. While MacDonald (1999), Race and MacDonald (2003), and others have also shown that multiple cognitive processes—including online comprehension in addition to production and offline judgment—draw on like probabilistic distributions, clearly much remains to be explored in teasing out the differences among these complex cognitive functions.
More narrowly, elements of experimental design could certainly further affect the ratings used to establish correlation. While the number of observations involved in building this study’s judgment models was more than sufficient to avoid overfitting, given the number of parameters included, the volume here (for example, N=729 after exclusions in the initial CC 100-point study) is of course nearly an order of magnitude smaller than that employed in building the corpus model (N=6,716).

Next, the stimuli in such controlled experiments are quite limited—just 30 variants in each trial—as compared to the natural speech of the Switchboard corpus from which the production model was built. It is also worth noting that this study’s design employed no fillers. This is in line with the similar experiment of syntactic alternation in Bresnan 2007, but it means that participants were likely to become quite conscious of the task. Next, this task obviously involved reviewing the stimuli in written, not spoken, form, and further, the judgment—as prompted by the task—required some degree of imagination on the part of the participant. Participants were asked for a rating of which alternative was likely to have been the actual continuation of speech, where such probabilistic anticipatory prediction may or may not be part of normal linguistic processing in conversation. Finally, while some prior discourse context was provided with each stimulus, it was of course only a snippet—only up to a few previous sentences—certainly far short of the entire rich pragmatic context on which language users can normally draw. In fact, in light of these limitations, it is substantial that participants make as accurate predictions as they do.

5. Conclusion

The present study comprises just one portion of a broader research program investigating Uniform Information Density as a potentially important cognitive force in linguistic processing. Other more recent work in this area includes supporting evidence both from other English language constructions—including contractions (Frank & Jaeger 2008) and hybrid phonological/syntactic reduction in the want to / wanna phenomenon (Melnick, in progress), among others—as well as cross-linguistically (cf. Kurumada 2011; inter alia). These and other works seek to continue to study the effect of UID on judgment and what this may reveal about the nature and content of linguistic competence.

6. References


## Appendix A: Complement-clause stimuli

<table>
<thead>
<tr>
<th>Token</th>
<th>Context</th>
<th>Stimulus</th>
</tr>
</thead>
<tbody>
<tr>
<td>practice 1</td>
<td>S1: Okay I've been there. Um, one down there is the uh...</td>
<td>I think (that) it's an old schoolhouse.</td>
</tr>
<tr>
<td>practice 2</td>
<td>S1: Uh huh, I agree with that. S2:</td>
<td>Uh, when I went to school I know (that) I was scared to death about being disciplined.</td>
</tr>
<tr>
<td>practice 3</td>
<td>S1: I'm uh interested in the shortwave aspect of it.</td>
<td>Uh, how do you find (that) the coverage from other places in the world uh compares with the American coverage?</td>
</tr>
<tr>
<td>1</td>
<td>S1: Yeah, painting, lawn, gutters, you name it. All kinds of stuff, flower beds. S2:</td>
<td>I think (that) that's considered exercise whether you want it or not.</td>
</tr>
<tr>
<td>2</td>
<td>S1: Well, are you going to paint the outside of your house too? S2:</td>
<td>I think (that) I am going to do it this spring actually.</td>
</tr>
<tr>
<td>3</td>
<td>S1: Um, Bonnie Rait, I like her a lot. Um, in fact, whenever she won her big Grammy...</td>
<td>I think (that) it was like five of them or something.</td>
</tr>
<tr>
<td>4</td>
<td>S1: I'm, uh, not for gun control, in the strictest sense of the, uh, word. S2:</td>
<td>I, I think (that) I'm kind of bent towards middle liberal of the bridge myself.</td>
</tr>
<tr>
<td>5</td>
<td>S1: Are you a player? S2:</td>
<td>Uh, I, I think (that) I am.</td>
</tr>
<tr>
<td>6</td>
<td>S1: I didn’t, wasn’t expecting that, so... S2: Well, that sounds great. S1:</td>
<td>I guess (that) in my spare time I’ll be making T shirts.</td>
</tr>
<tr>
<td>7</td>
<td>S1: Do you do exercise? S2: I do, yes.</td>
<td>I, uh, uh, I guess (that) it actually changes.</td>
</tr>
<tr>
<td>8</td>
<td>S1: It’s a scream, but I have to get up and, and work the next morning. S2:</td>
<td>I wish (that) they’d put those on.</td>
</tr>
<tr>
<td>9</td>
<td>S1: Uh, the State of Wisconsin, as a matter of fact, uh, started some litigation against Illinois because of the air pollution we were getting. S2: Uh huh. S1:</td>
<td>Uh, I don’t think (that) it’s going to go very far.</td>
</tr>
<tr>
<td>10</td>
<td>S1: But it is, uh, it is bad everywhere in terms of, uh, you know, the handgun situation, um, the number of rapes, the number of muggings.</td>
<td>Just, it is, it is, I would not say (that) it is not as bad as New York.</td>
</tr>
<tr>
<td>11</td>
<td>S1: Uh, and it’s, it’s kind of small in a sense, but it, it does have about, uh, I guess two thousand civil employees. S2: Uh huh. S1:</td>
<td>And, on a scale, I guess (that) it’s still considered small.</td>
</tr>
<tr>
<td>12</td>
<td>S1: I do not know whether the jury system, uh, ...</td>
<td>I, I should not say (that) I do not know.</td>
</tr>
<tr>
<td>13</td>
<td>S1: Uh, in Dallas there’s definitely a crime problem.</td>
<td>I know (that) it’s, you know, worse in the larger cities.</td>
</tr>
<tr>
<td>14</td>
<td>S1: I have a unique situation here in that if I don’t participate... S2:</td>
<td>And you say (that) you have some strong feelings about, uh, the issue?</td>
</tr>
<tr>
<td>15</td>
<td>S1: Well, I’m real curious because my family, it didn’t sound...</td>
<td>I don’t think (that) my family is as big as your husband’s.</td>
</tr>
<tr>
<td>16</td>
<td>S1: Well, I’m kind of in favor of it for certain crimes. S2: Yeah. Which crimes do you feel that... S1:</td>
<td>Well, I think (that) first degree murder, uh, probably warrants it.</td>
</tr>
<tr>
<td></td>
<td>S1: Well, we've always, uh, we've always had Oldsmobiles, and, uh, been very, uh, happy with Oldsmobiles...</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------------------------</td>
<td>---</td>
</tr>
<tr>
<td>17</td>
<td>S1: And what is your feeling?</td>
<td>18</td>
</tr>
<tr>
<td>18</td>
<td>S1: Uh, but that depends, you know, on the individual if they can I guess have self-control. S2: Well, that's a lot of it.</td>
<td>19</td>
</tr>
<tr>
<td>19</td>
<td>S1: I'm concerned about potential for injury to your knees, you know, and movement of going up the step. Could create a problem. S2: Oh, yeah. S1:</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>S1: I like, I like the older ones because at that time, you know, they were really futuristic, you know. S2: Oh, yeah. S2:</td>
<td>21</td>
</tr>
<tr>
<td>21</td>
<td>S1: How do you feel about gun control? S2:</td>
<td>22</td>
</tr>
<tr>
<td>22</td>
<td>S1: How would you feel about sending an elderly family member to a nursing home? S2: Well, I don't feel very good about it...</td>
<td>23</td>
</tr>
<tr>
<td>23</td>
<td>S1: He might, could get hurt by a car or something.</td>
<td>24</td>
</tr>
<tr>
<td>24</td>
<td>S1: I wonder how truthful all of that was or whether there was fiction. S2: Yeah.</td>
<td>25</td>
</tr>
<tr>
<td>25</td>
<td>S1: I watched a lot of CNN because it was so good. S2:</td>
<td>26</td>
</tr>
<tr>
<td>26</td>
<td>S1: But, uh, you know, the transmission may be made in Japan or whatever. Uh, like I've got an eighty-six Ford Ranger...</td>
<td>27</td>
</tr>
<tr>
<td>27</td>
<td>S1: How did you ever get into that? That sounds so interesting. S2: My husband kept begging me, and he's been dead now for twenty years...</td>
<td>28</td>
</tr>
<tr>
<td>28</td>
<td>S1: The longevity of the house is not, uh, is not worth it. How about in your case? S2: Well, in my case my husband is not a carpenter, but in fact, he's in electronics...</td>
<td>29</td>
</tr>
<tr>
<td>29</td>
<td>S1: Well, it is fun, what little I do.</td>
<td>30</td>
</tr>
</tbody>
</table>
### Appendix B: Complement-clause mean judgment ratings summary

<table>
<thead>
<tr>
<th>Token</th>
<th>Verb (lemma)</th>
<th>Information Density</th>
<th>Corpus fitted</th>
<th>Mean ratings</th>
<th>Lab</th>
<th>Crowdsourced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>100-points</td>
<td>forced-choice</td>
<td>100-points</td>
<td>forced-choice</td>
</tr>
<tr>
<td>1</td>
<td>think</td>
<td>0.670</td>
<td>0.019</td>
<td>0.210</td>
<td>0.179</td>
<td>0.285</td>
</tr>
<tr>
<td>2</td>
<td>think</td>
<td>0.670</td>
<td>0.021</td>
<td>0.361</td>
<td>0.179</td>
<td>0.379</td>
</tr>
<tr>
<td>3</td>
<td>think</td>
<td>0.670</td>
<td>0.024</td>
<td>0.296</td>
<td>0.379</td>
<td>0.400</td>
</tr>
<tr>
<td>4</td>
<td>think</td>
<td>0.670</td>
<td>0.034</td>
<td>0.384</td>
<td>0.448</td>
<td>0.584</td>
</tr>
<tr>
<td>5</td>
<td>think</td>
<td>0.670</td>
<td>0.042</td>
<td>0.196</td>
<td>0.107</td>
<td>0.358</td>
</tr>
<tr>
<td>6</td>
<td>guess</td>
<td>0.623</td>
<td>0.005</td>
<td>0.360</td>
<td>0.143</td>
<td>0.418</td>
</tr>
<tr>
<td>7</td>
<td>guess</td>
<td>0.623</td>
<td>0.005</td>
<td>0.362</td>
<td>0.385</td>
<td>0.367</td>
</tr>
<tr>
<td>8</td>
<td>wish</td>
<td>0.904</td>
<td>0.050</td>
<td>0.251</td>
<td>0.241</td>
<td>0.414</td>
</tr>
<tr>
<td>9</td>
<td>think</td>
<td>0.670</td>
<td>0.075</td>
<td>0.408</td>
<td>0.321</td>
<td>0.433</td>
</tr>
<tr>
<td>10</td>
<td>say</td>
<td>0.266</td>
<td>0.108</td>
<td>0.524</td>
<td>0.423</td>
<td>0.546</td>
</tr>
<tr>
<td>11</td>
<td>guess</td>
<td>0.623</td>
<td>0.009</td>
<td>0.365</td>
<td>0.345</td>
<td>0.515</td>
</tr>
<tr>
<td>12</td>
<td>say</td>
<td>0.266</td>
<td>0.134</td>
<td>0.569</td>
<td>0.357</td>
<td>0.587</td>
</tr>
<tr>
<td>13</td>
<td>know</td>
<td>0.048</td>
<td>0.081</td>
<td>0.480</td>
<td>0.536</td>
<td>0.566</td>
</tr>
<tr>
<td>14</td>
<td>say</td>
<td>0.266</td>
<td>0.089</td>
<td>0.494</td>
<td>0.630</td>
<td>0.696</td>
</tr>
<tr>
<td>15</td>
<td>think</td>
<td>0.670</td>
<td>0.122</td>
<td>0.416</td>
<td>0.214</td>
<td>0.508</td>
</tr>
<tr>
<td>16</td>
<td>think</td>
<td>0.670</td>
<td>0.110</td>
<td>0.576</td>
<td>0.571</td>
<td>0.584</td>
</tr>
<tr>
<td>17</td>
<td>think</td>
<td>0.670</td>
<td>0.180</td>
<td>0.481</td>
<td>0.429</td>
<td>0.469</td>
</tr>
<tr>
<td>18</td>
<td>think</td>
<td>0.670</td>
<td>0.092</td>
<td>0.389</td>
<td>0.500</td>
<td>0.605</td>
</tr>
<tr>
<td>19</td>
<td>know</td>
<td>0.048</td>
<td>0.146</td>
<td>0.409</td>
<td>0.440</td>
<td>0.595</td>
</tr>
<tr>
<td>20</td>
<td>think</td>
<td>0.670</td>
<td>0.228</td>
<td>0.523</td>
<td>0.391</td>
<td>0.590</td>
</tr>
<tr>
<td>21</td>
<td>guess</td>
<td>0.623</td>
<td>0.020</td>
<td>0.287</td>
<td>0.179</td>
<td>0.417</td>
</tr>
<tr>
<td>22</td>
<td>think</td>
<td>0.670</td>
<td>0.203</td>
<td>0.533</td>
<td>0.464</td>
<td>0.582</td>
</tr>
<tr>
<td>23</td>
<td>think</td>
<td>0.670</td>
<td>0.189</td>
<td>0.457</td>
<td>0.679</td>
<td>0.492</td>
</tr>
<tr>
<td>24</td>
<td>know</td>
<td>0.048</td>
<td>0.434</td>
<td>0.703</td>
<td>0.643</td>
<td>0.675</td>
</tr>
<tr>
<td>25</td>
<td>imagine</td>
<td>0.304</td>
<td>0.198</td>
<td>0.509</td>
<td>0.429</td>
<td>0.518</td>
</tr>
<tr>
<td>26</td>
<td>think</td>
<td>0.670</td>
<td>0.161</td>
<td>0.490</td>
<td>0.393</td>
<td>0.569</td>
</tr>
<tr>
<td>27</td>
<td>know</td>
<td>0.048</td>
<td>0.504</td>
<td>0.498</td>
<td>0.720</td>
<td>0.662</td>
</tr>
<tr>
<td>28</td>
<td>understand</td>
<td>0.088</td>
<td>0.545</td>
<td>0.542</td>
<td>0.630</td>
<td>0.526</td>
</tr>
<tr>
<td>29</td>
<td>know</td>
<td>0.048</td>
<td>0.678</td>
<td>0.511</td>
<td>0.643</td>
<td>0.514</td>
</tr>
<tr>
<td>30</td>
<td>tell</td>
<td>0.210</td>
<td>0.541</td>
<td>0.536</td>
<td>0.593</td>
<td>0.645</td>
</tr>
</tbody>
</table>
## Appendix C: Relative-clause stimuli

<table>
<thead>
<tr>
<th>Token</th>
<th>Context</th>
<th>Stimulus</th>
</tr>
</thead>
<tbody>
<tr>
<td>practice 1</td>
<td>S1: I know in baseball, you know any manager says, well, we're going to be in the World Series, even though you know there's not a prayer. S2: Yeah, really. S1:</td>
<td>this is the way it is</td>
</tr>
<tr>
<td>practice 2</td>
<td>S1: I've never asked her any questions, but she has three kids all under the age of like six. And, and they're all a handful.</td>
<td>so it's got to be something she can do fairly easily and fairly quickly</td>
</tr>
<tr>
<td>practice 3</td>
<td>S1: But, even if they had three-car garages, they didn't have a work space. All of it was taken up with just car space. S2: Uh-huh. S1:</td>
<td>and so he has a big landing area that's between where the cars are in the garage and where the house starts that he can work on</td>
</tr>
<tr>
<td>1</td>
<td>S1: There's been a couple of other people, most of whom are like me that work in, in speech labs that are going to use database, eventually, so, it was fun.</td>
<td>So the day (that) it happens is ninety-eight percent of the planning usually.</td>
</tr>
<tr>
<td>2</td>
<td>S1: And they still keep asking you, you know, and keep badgering you about it, and that aggravates me. But I don't know that you could call that invading of, invading my privacy. Because, you know, if we don't want that to happen ...</td>
<td>I think a lot of it is the families, the way (that) you were mentioning.</td>
</tr>
<tr>
<td>3</td>
<td>S1: And what did, did you ever try using like Prodigy or any of those systems? S2: No, I haven't done that. I know someone who has, and, and she's very pleased with it.</td>
<td>And it's, it's everything (that) it's supposed to be</td>
</tr>
<tr>
<td>4</td>
<td>S1: And finally, you know, who knows, maybe they're finally waking up and saying, you know, we can't afford this. S2:</td>
<td>So next thing (that) you know she's teaching in a maximum security prison.</td>
</tr>
<tr>
<td>5</td>
<td>S1: Oh, that's good, that's good, or it's always good on top of something in the sandwich. S2: Yeah, yeah. S1:</td>
<td>So hers was the first one (that) I got.</td>
</tr>
<tr>
<td>6</td>
<td>S1: I go home, I've gone home every year now since I've moved to Dallas to, to go pheasant hunting, and the, the last time I took my wife along, and kind of the same situation.</td>
<td>Al Jarreau, he's somebody (that) I like.</td>
</tr>
<tr>
<td>7</td>
<td>S1: I, I don't think that they, they educate them enough to, to really know what's going on. S2:</td>
<td>She told me some of the things (that) you can do.</td>
</tr>
<tr>
<td>8</td>
<td>S1: That's on my list to see. S2: Yeah, it was, it was okay. It was not, not a wonderful film ...</td>
<td>I can remember the first year (that) I moved here.</td>
</tr>
</tbody>
</table>
S1: Oh, I think we rented 'Pretty Woman' a couple weeks, months ago. It requires a lot of effort to, to do that sort of thing. I mean, I used to enjoy going, I mean I still do kind of enjoy it ...

Because, you know, if we don't want that to happen, all (that) we have to do is just call the phone company and say look, you know, i want my name unlisted, or want my, you know.

S1: You know, sometimes it's just so, seems so much easier just to take it and throw it in the trash, and have them pick it up than it is to smash the cans and drive it someplace to have them ...

And there's only so much football (that) you can watch.

S1: Yeah, I know, that's kind of how I am. When it's summertime, I'm wearing shorts, ...

She's up in the house and talked to my mother the whole time (that) we were out hunting and stuff.

S1: And we take the, the civil liberties stuff too far. You know like people that are in prison, I mean we didn't put them there. They put themselves in that situation.

I mean I don't, I honestly don't really think I could have, you know, done much better than I did in the school system (that) I was in.

S2: Uh-huh, oh yeah. S1: But it's not as many people there that I really, really enjoy seeing.

S1: I don't really bother with the Washington station because I just, it's so far removed from what I'm interested in.

And the rates that we keep paying seem to keep increasing, you know.

S1: Well, I think, the, the health care, of course, is an important, and has to be the single most important benefit.

In the case (that) you were involved in, you said it was just sort of a bank matter of some kind.

S1: Well, the summers have gotten where they're a lot hotter.

Actually yesterday I ended up talking to somebody else from the same lab (that) I'm in.

S1: I think they checked for drugs. They just don't tell you they did.

But it was, I mean, it was, it was cute, but not the biggest laugh (that) I've seen.

S1: Do y'all start planning real far ahead of time? S2: Well, no. There's kind of a set pattern to it. There is, ...

Well, most of what they're talking about in Washington is is the crime problems that they're having there.

S1: Yeah, yeah, I can, I can believe that. Yeah, I, I don't claim to have an in-depth understanding by any means, but ... S2: Now I, shoot, ...

The U.S. isn't the threat that we've always made them out to be, you know, even if they're saying that beneath the doors.

S1: But, well, she had decided that it was, when we came back to the States, she decided that it was too, she didn't feel that public schools were, were safe.

See I went out to Payless Cashways here a couple of days ago as a matter of fact and got me one of those little can crushers that I could put on the wall.

S1: I love them. I have some dulcimer music. I mean, like I said, that, that's pretty varied. S2: Yeah, it sure is. S1: They convicted him of a robbery that he could not physically have been able to commit.

S1: Absolutely. I have had it for about two years now ... S2: I made guacamole that I would take to a pool party across the street the other day.
<table>
<thead>
<tr>
<th>S1:</th>
<th>S2:</th>
<th>S1:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well I, well I work for the government, and, actually I work for the FBI.</td>
<td>Oh, my gosh.</td>
<td>But the the ice storms really are stressful on all the, the plants that we have around.</td>
</tr>
<tr>
<td>I would put just a little bit of cat litter in there, because if I put a lot, one of the cats would have been, ...</td>
<td>Uh-huh.</td>
<td>I, I think some of these people that they claim as middle eastern experts don't, don't have a clue sometimes.</td>
</tr>
<tr>
<td>And then tonight this woman called. Have you taken any incoming calls?</td>
<td>Yours is my second one.</td>
<td>And when it's cold I'm wearing sweatpants unless I'm going to something special that my daughter's doing or something like that.</td>
</tr>
<tr>
<td>You get those great big cracks in the ground where it's been so dry, and it just gets so dry, and the earth opens up.</td>
<td></td>
<td>There's lots of things that we don't get told for good reason.</td>
</tr>
<tr>
<td>And the judicial mishap in this event would be, hey, you got the wrong guy. Look at that guy at E Systems.</td>
<td>Yep.</td>
<td>I have a few favorites that I use more than others.</td>
</tr>
<tr>
<td>Yeah, they seem to be a part of life. Do you use them?</td>
<td>Well, I do use them.</td>
<td>All of her animals that she ever had were adopted.</td>
</tr>
<tr>
<td>If you compare the, like the people that could have, that could have sent their kids to private schools and the people that, you know, the people that did send their kids to private schools, I think they compare, you know, fairly well.</td>
<td></td>
<td>I had, I've had two or three drug tests that I had to get before I could start working at a job, different jobs.</td>
</tr>
<tr>
<td>And you're, you get mixed signals. So it's, it's not always the school systems.</td>
<td></td>
<td>And as far as, like, them entertaining the rights that they should have.</td>
</tr>
</tbody>
</table>
## Appendix D: Relative-clause mean judgment ratings summary

<table>
<thead>
<tr>
<th>Token</th>
<th>Head noun (lemma)</th>
<th>Corpus fitted</th>
<th>100-points</th>
<th>forced-choice</th>
<th>100-points</th>
<th>forced-choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>day</td>
<td>0.055</td>
<td>0.407</td>
<td>0.240</td>
<td>0.471</td>
<td>0.360</td>
</tr>
<tr>
<td>2</td>
<td>way</td>
<td>0.150</td>
<td>0.365</td>
<td>0.148</td>
<td>0.416</td>
<td>0.292</td>
</tr>
<tr>
<td>3</td>
<td>everything</td>
<td>0.089</td>
<td>0.491</td>
<td>0.500</td>
<td>0.472</td>
<td>0.542</td>
</tr>
<tr>
<td>4</td>
<td>thing</td>
<td>0.114</td>
<td>0.295</td>
<td>0.179</td>
<td>0.313</td>
<td>0.379</td>
</tr>
<tr>
<td>5</td>
<td>one</td>
<td>0.085</td>
<td>0.387</td>
<td>0.458</td>
<td>0.414</td>
<td>0.481</td>
</tr>
<tr>
<td>6</td>
<td>somebody</td>
<td>0.145</td>
<td>0.408</td>
<td>0.261</td>
<td>0.476</td>
<td>0.273</td>
</tr>
<tr>
<td>7</td>
<td>thing</td>
<td>0.278</td>
<td>0.499</td>
<td>0.321</td>
<td>0.405</td>
<td>0.500</td>
</tr>
<tr>
<td>8</td>
<td>year</td>
<td>0.289</td>
<td>0.370</td>
<td>0.320</td>
<td>0.410</td>
<td>0.240</td>
</tr>
<tr>
<td>9</td>
<td>all</td>
<td>0.208</td>
<td>0.362</td>
<td>0.214</td>
<td>0.323</td>
<td>0.346</td>
</tr>
<tr>
<td>10</td>
<td>football</td>
<td>0.331</td>
<td>0.413</td>
<td>0.259</td>
<td>0.325</td>
<td>0.364</td>
</tr>
<tr>
<td>11</td>
<td>time</td>
<td>0.381</td>
<td>0.402</td>
<td>0.393</td>
<td>0.400</td>
<td>0.480</td>
</tr>
<tr>
<td>12</td>
<td>system</td>
<td>0.358</td>
<td>0.476</td>
<td>0.333</td>
<td>0.558</td>
<td>0.560</td>
</tr>
<tr>
<td>13</td>
<td>people</td>
<td>0.533</td>
<td>0.558</td>
<td>0.440</td>
<td>0.653</td>
<td>0.792</td>
</tr>
<tr>
<td>14</td>
<td>rate</td>
<td>0.512</td>
<td>0.387</td>
<td>0.360</td>
<td>0.508</td>
<td>0.333</td>
</tr>
<tr>
<td>15</td>
<td>case</td>
<td>0.452</td>
<td>0.509</td>
<td>0.360</td>
<td>0.539</td>
<td>0.625</td>
</tr>
<tr>
<td>16</td>
<td>lab</td>
<td>0.415</td>
<td>0.455</td>
<td>0.407</td>
<td>0.525</td>
<td>0.615</td>
</tr>
<tr>
<td>17</td>
<td>laugh</td>
<td>0.457</td>
<td>0.367</td>
<td>0.357</td>
<td>0.506</td>
<td>0.519</td>
</tr>
<tr>
<td>18</td>
<td>problem</td>
<td>0.532</td>
<td>0.535</td>
<td>0.385</td>
<td>0.459</td>
<td>0.519</td>
</tr>
<tr>
<td>19</td>
<td>threat</td>
<td>0.667</td>
<td>0.500</td>
<td>0.407</td>
<td>0.483</td>
<td>0.440</td>
</tr>
<tr>
<td>20</td>
<td>crusher</td>
<td>0.775</td>
<td>0.437</td>
<td>0.560</td>
<td>0.605</td>
<td>0.519</td>
</tr>
<tr>
<td>21</td>
<td>robbery</td>
<td>0.789</td>
<td>0.518</td>
<td>0.571</td>
<td>0.643</td>
<td>0.600</td>
</tr>
<tr>
<td>22</td>
<td>guacamole</td>
<td>0.784</td>
<td>0.577</td>
<td>0.607</td>
<td>0.657</td>
<td>0.704</td>
</tr>
<tr>
<td>23</td>
<td>plant</td>
<td>0.613</td>
<td>0.415</td>
<td>0.320</td>
<td>0.433</td>
<td>0.375</td>
</tr>
<tr>
<td>24</td>
<td>people</td>
<td>0.698</td>
<td>0.499</td>
<td>0.556</td>
<td>0.471</td>
<td>0.417</td>
</tr>
<tr>
<td>25</td>
<td>something</td>
<td>0.865</td>
<td>0.576</td>
<td>0.600</td>
<td>0.506</td>
<td>0.680</td>
</tr>
<tr>
<td>26</td>
<td>lot</td>
<td>0.800</td>
<td>0.463</td>
<td>0.688</td>
<td>0.503</td>
<td>0.478</td>
</tr>
<tr>
<td>27</td>
<td>favorite</td>
<td>0.853</td>
<td>0.528</td>
<td>0.577</td>
<td>0.591</td>
<td>0.607</td>
</tr>
<tr>
<td>28</td>
<td>all</td>
<td>0.936</td>
<td>0.529</td>
<td>0.391</td>
<td>0.473</td>
<td>0.227</td>
</tr>
<tr>
<td>29</td>
<td>test</td>
<td>0.923</td>
<td>0.429</td>
<td>0.607</td>
<td>0.575</td>
<td>0.600</td>
</tr>
<tr>
<td>30</td>
<td>right</td>
<td>0.837</td>
<td>0.507</td>
<td>0.481</td>
<td>0.440</td>
<td>0.458</td>
</tr>
</tbody>
</table>
Appendix E: Regression model predictors

Independent variables from the prior corpus model (Jaeger 2010), excluding those factors omitted from the present study as unavailable to offline judgment participants (original speech rate, speaker gender, and so on; See Table 1).

All information here is sourced from Jaeger 2010, pps. 30-36.

Random variables

• SPEAKER ID of current speaker (corpus model) or participant (judgment/experiment model), to control for per-speaker that-mentioning bias

• LEMMA(MATRIX VERB) Lemmatized form of matrix verb, to control for per-verb bias for that-mentioning independent of the verb’s particular predictability of clausal embedding (the latter being the information density / UID predictor)

Fixed variables

• INFORMATION(CC ONSET) The primary measure of interest, reflects information density as the -log conditional probability of CC onset in context—specifically estimated as the observed likelihood of clausal embedding for the given matrix verb in the Switchboard corpus. [continuous]

• POSITION(MATRIX VERB) Position in the sentence of the matrix (embedding) verb, modeled as a restricted cubic spline. [3x continuous]

• LENGTH(MATRIX VERB-TO-CC) Distance (in words) from matrix (embedding) verb to CC onset (including that if mentioned). [continuous]

• LENGTH(CC ONSET) Number of words up to and including the CC subject. [continuous]

• LENGTH(CC REMAINDER) Number of words in the CC beyond its subject. [continuous]
• CC SUBJECT
  Four-way categorical encoding of type of referential expression represented by the CC subject—*I, it, other pronoun, other NP*—yielding three contrast-encoded predictors against baseline value *I*. [3x categorical]

• SUBJECT IDENTITY
  Binary encoding of whether or not the CC subject is string-identical to the matrix subject. [categorical]

• FREQUENCY(CC SUBJECT HEAD)
  Log-transformed corpus frequency of lemmatized CC subject head noun. [continuous]

• WORD FORM SIMILARITY
  Binary encoding for whether or not the CC begins with demonstrative *that*; reflects the potential for a double-*that* sequence. [categorical]

• FREQUENCY(MATRIX VERB)
  Log corpus frequency of lemmatized matrix (embedding) verb. [continuous]

• AMBIGUOUS CC ONSET
  Binary encoding for potential garden-path ambiguity at CC onset. Nominative-case CC subjects and tokens with matrix verbs that never or rarely take a NP complement are encoded as unambiguous. Other tokens evaluated by hand. [categorical]

• MATRIX SUBJECT
  Four-way categorical encoding of type of referential expression represented by the matrix subject—*I, you, other pronoun, other NP*—yielding three treatment-coded predictors against baseline value *I*. [3x categorical]