Uniform Information Density and Syntactic Reduction: Revisiting *that*-Mentioning in English Complement Clauses

Anonymous ACL submission

Abstract

Speakers often have multiple ways to express the same meaning. The Uniform Information Density (UID) hypothesis suggests that speakers exploit this variability to maintain a consistent rate of information transmission during language production. Building on prior work linking UID to syntactic reduction, we revisit the finding that the optional complementizer *that* in English complement clauses is more likely to be omitted when the clause has low information density (i.e., more predictable). We advance this line of research by analyzing a large-scale, contemporary conversational corpus and using machine learning and neural language models to refine estimates of information density. Our results replicated the established relationship between information density and *that*-mentioning. However, we found that previous measures of information density based on matrix verbs' subcategorization probability capture substantial idiosyncratic lexical variation. By contrast, estimates derived from contextual word embeddings account for additional variance in patterns of complementizer usage.[1]

1 Introduction

Language production is highly flexible across all levels of linguistic analysis, such as phonetics, lexicon, and syntax. Such flexibility in production enables researchers to ask the question: What cognitive mechanisms guide our choice among competing alternatives? A prominent account, Uniform Information Density (UID; Jaeger, 2010; Levy and Jaeger, 2007), proposes that speakers exploit this flexibility to maintain a consistent rate of information transmission. According to UID, speakers tend to structure their utterances to distribute information as evenly as possible across the linguistic signal to ensure robust information transmission while maintaining efficient use of the communication channel. Following Shannon's (1948) information theory, the information density of a unit u given its context is defined as:

$$I(u) = -\log(P(u \mid \text{context})) \quad (1)$$

where $P(u \mid \text{context})$ denotes the contextual probability of u. This is also commonly referred to as surprisal (Halle, 2001; Levy, 2008).

In an influential study, Jaeger (2010) demonstrated that UID effects can be observed at the syntactic level. He examined the optional complementizer *that* in English complement clauses (henceafter CCs; e.g., (1)) and found that *that* is more likely to be included when the information density of the CC is high—that is, when the contextual probability of a CC given the preceding context, P(CC|context), is low.

(1) The boss complained (that) they were crazy.

The rationale is that an unpredicted CC would create a spike in information density at the clause onset without *that*, since the CC is unexpected, while including *that* helps smooth the distribution by signaling the upcoming structure. Conversely, when a CC is highly predictable, *that* becomes redundant and may introduce an information density trough. This preference is illustrated in Figure 1. When the CC onset is information-heavy, potentially exceeding the channel's capacity (Figure 1a), including *that* can reduce peak information density (Figure 1b). In contrast, when the onset is relatively low in information density, mentioning *that* would create a valley (Figure 1c), while omitting it results in a smoother information profile (Figure 1d).

While Jaeger (2010) provided important evidence for UID, several limitations remain in this work. First, P(CC|context) was quantified using matrix verbs' subcategorization probabilities—that is, the proportion of times a given verb (based on its lemma) takes a CC as its syntactic object, based

[1] Code is available anonymously here.

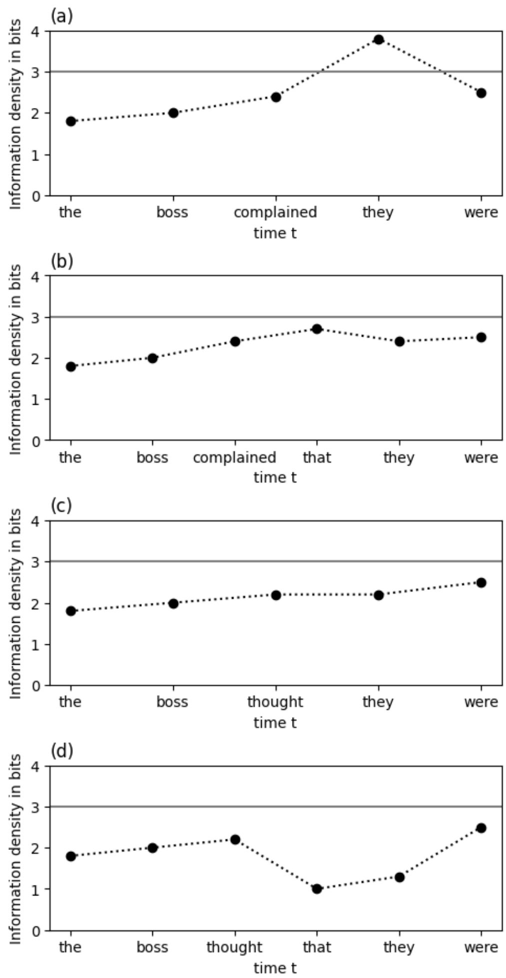

Figure 1: Illustration of per-word information density (gray line represents a hypothetical channel capacity): (a) High information density at CC onset; (b) Including *that* to reduce peak information; (c) Low information density at onset with *that*-mentioning; (d) Smooth profile without *that*.

on corpus-derived frequencies. This static measure might not not fully capture dynamic predictive processes and may conflate predictability with verb-specific variation. Second, the study used a relatively small and outdated dataset: about 8,000 CCs with 31 matrix verbs from the Switchboard corpus (Godfrey et al., 1992; Marcus et al., 1999), which may limit the generalizability of the findings. Given the theoretical importance of Jaeger's (2010) findings, a reexamination using larger datasets and more refined predictability measures is needed.

To address these limitations, in the current work we analyzed a modern large-scale corpus called Conversation: A Naturalistic Dataset of Online Recordings (CANDOR; Reece et al., 2023). To preview, we extracted over 50,000 unique cases of CCs after data cleaning, encompassing 50 unique matrix verbs. In addition, we incorporated insights from machine learning and neural language models, especially contextual word embeddings, to refine measures of structural predictability. Such refined estimation also allows us to investigate whether improved predictability of CCs leads to better modeling of *that*-mentioning.

2 Related Work

2.1 Psycholinguistic Evidence for UID

Research supporting the UID hypothesis in language production spans multiple linguistic levels, including phonetics (Aylett and Turk, 2004), lexical choice (Mahowald et al., 2013), syntax (Jaeger, 2010), and discourse (Asr and Demberg, 2015). For example, past research across many different languages has consistently demonstrated that when a word or phoneme is more predictable in context, it is typically produced with a shorter duration and exhibits reduced phonological and phonetic detail (Aylett and Turk, 2004; Bell et al., 2009; Cohen Priva, 2015; Pimentel et al., 2021; Pluymaekers et al., 2005, among others). At the lexical level, Mahowald et al. (2013) found that speakers are more likely to use shortened forms of words (e.g., math instead of mathematics) in more predictive contexts. Similarly, at the syntactic level, studies have shown that optional syntactic markers, such as *that* in English CCs (e.g., *I think (that) the weather is very nice*; Jaeger, 2010) and object relative clauses (e.g., *the groceries (that) they brought home*; Levy and Jaeger, 2007), are more frequently omitted when the upcoming syntactic structure is highly predictable. The relationship between information density and syntactic reduction also extends cross-linguistically, such as in subject doubling in French (Liang et al., 2024) and optional indefinite articles in German (Lemke et al., 2017).

However, the predictions of UID are not always borne out. For example, Zhan and Levy (2018) found that variation in the use of specific versus general classifiers in Mandarin Chinese is better explained by availability-based production accounts. In addition, Kuperman et al. (2007) observed that Dutch interfixes are pronounced longer when they have higher contextual probability, contrary to UID predictions, which they attributed to paradigmatic enhancement. These divergent findings underscore the need for further evaluation of UID.

2.2 Neural Language Model and Structural Knowledge

A range of studies has probed neural language models' sensitivity to linguistic structures. Linzen et al. (2016), for instance, evaluated LSTMs' ability to capture subject-verb agreement using template-based test data. Extending this approach, Warstadt et al. (2020) developed a broader benchmark encompassing a diverse set of linguistic phenomena (see also Hu et al., 2020). Many of these studies rely on surprisal-based evaluations, assuming that ungrammatical continuations should elicit higher surprisal than grammatical ones (e.g., Futrell et al., 2019; Wilcox et al., 2018). Other work has adapted stimuli from psycholinguistic experiments, comparing language model surprisal to human behavioral or neural data (Arehalli and Linzen, 2020; Hao, 2023; Huang et al., 2024; Michaelov and Bergen, 2020). For critical overviews of this literature, see Limisiewicz and Mareček (2020) and Linzen and Baroni (2021)

Beyond surprisal-based evaluations, researchers have also assessed models' syntactic knowledge through attention head analyses (e.g., Clark et al., 2019; Ryu and Lewis, 2021), meta-linguistic prompting (e.g., Dentella et al., 2024; Katzir, 2023; Zhou et al., 2023; though see Hu and Levy, 2023, for critiques of this method), and examinations of contextual word embeddings (e.g., Li et al., 2022; Peters et al., 2018; Petty et al., 2022; Tenney et al., 2019; Wilson et al., 2023). For instance, Peters et al. (2018) demonstrated that contextual embeddings encode a wide range of syntactic information, such as part-of-speech and syntactic boundaries, while Li et al. (2022) showed that contextual word embeddings are sensitive to argument structure even in semantically anomalous sentences.

3 Structural Predictability Model

We trained several neural binary classifiers using either hand-selected linguistic features from the pre-CC context or contextual word embeddings of the matrix verb, to estimate the structural predictability of CCs. Hand-selected features offer interpretability and theoretical grounding but may overlook subtle or high-dimensional patterns in the linguistic context. In contrast, contextual word embeddings (e.g., from BERT or GPT models) encode nuanced semantic and syntactic information by capturing how the meaning of a word dynamically changes depending on its surrounding context, but come at the cost of interpretability (Kennedy et al., 2021). To balance the trade-off, we evaluate models trained on each feature type separately.

3.1 Linguistic Features

We included features from the matrix verb and the matrix subject in the pre-CC context. For the matrix verb, we included its subcategorization probability, estimated from the CANDOR corpus (see Appendix A), as well as its log frequency (SUBTLEX; Brysbaert and New, 2009), factivity (i.e., whether it presupposes the truth of the clause it introduces); Karttunen, 1971), tense (base form vs. inflected), and position within the sentence. We also included two features related to the matrix subject: form (*I*, *You*, *Other pronouns* vs. *Other nouns*) and log frequency.

To identify the most effective feature set, we performed incremental feature selection, adding features one at a time starting from the matrix verb's subcategorization probability. A feature was retained only if it improved model fit according to Akaike Information Criteria (AIC; Akaike, 1974) and Bayesian Information Criteria (BIC; Schwarz, 1978), both of which balance model fit and complexity by penalizing the inclusion of unnecessary parameters. We also experimented with Lasso regression (Tibshirani, 1996), where we first fitted a linear regression model using all features simultaneously with an L1 penalty to encourage sparsity in the feature set. Features with nonzero coefficients were then used to predict CC presence.

3.2 Contextual Word Embeddings

To capture richer predictive cues, we extracted contextual embeddings of the matrix verb from GPT-2 Small (GPT-2 henceforth; Radford et al., 2019). Note that this context only includes pre-CC information, not information after the CC onset (e.g., we extracted the embeddings of *complained* from *the boss complained*). GPT-2's autoregressive architecture enables embeddings based solely on preceding context, aligning with incremental sentence processing. We used the final hidden state of the verb token and reduced the 768-dimensional embeddings to 50 dimensions via PCA (Jolliffe, 2002), preserving over 99% of the variance.

3.3 Training Data

The training data come from the CANDOR corpus (Reece et al., 2023), a large-scale dataset of 1,656 dyadic conversations recorded over Zoom. The corpus is publicly available and can be requested here. These conversations capture spontaneous, unscripted exchanges between strangers and are supplemented with detailed survey data. The corpus includes 1,456 unique participants representing a diverse range of gender identities, educational backgrounds, ethnicities, and generations. The mean conversation duration is 31.3 minutes (SD = 7.96, min = 20). All analyses in this study are based on existing transcripts from the corpus, totaling approximately 8 million words. Transcripts were segmented using the Cliffhanger algorithm, which groups utterances based on terminal punctuation (e.g., periods, exclamations, questions) and integrates backchannels into broader conversational units.

Transcripts were automatically parsed using spaCy's dependency parser (Honnibal et al., 2020), following Universal Dependencies conventions (de Marneffe et al., 2021; Nivre et al., 2016). We began with 86 matrix verbs that can take CCs, identified by Jaeger (2010) and Jaeger and Grimshaw (2013). Based on frequency in the CANDOR corpus, we selected the 50 most frequent verbs ($\geq$ 100 occurrences; see Appendix A).

We then extracted all instances of these 50 verbs, regardless of whether they were followed by a CC, direct object, or other dependents. We excluded cases where the verb was sentence-final or the matrix subject was missing. Each instance is labeled as **1** if followed by a CC and **0** otherwise. The final dataset consists of 236,504 training examples, with 33.01% labeled as **1**.

3.4 Model Architecture, Training, and Evaluation

We trained feedforward neural networks to predict CC presence. The input features are fed into three hidden layers (128, 64, and 32 units, respectively) with ReLU activation, batch normalization, and 0.2 dropout. The final layer uses sigmoid activation to produce probabilities ranging from 0 to 1. Before training, all numerical predictors are z-scored, and categorical variables factor-encoded.

The model is trained using binary cross-entropy loss and optimized with Adam (learning rate = 0.001, weight decay = 1e-5) in minibatches of 1024 instances. Training proceeds for up to 50 epochs, with early stopping if validation loss does not improve after five epochs. We used five-fold stratified cross-validation to maintain class distribution across splits.

3.5 Structural Predictability Model Results

Results from the incremental selection of linguistic features are presented in Table 1. Recall that a new feature was added only if it improved model performance in terms of AIC and BIC. Table 1 reports the change in AIC and BIC relative to the previously selected model. For reference, we also report each model's F1 score and log loss. As shown in Table 1, including subcategorization probability leads to reductions in both AIC and BIC relative to the baseline model, as well as lower log loss and higher F1 scores. However, none of the additional linguistic features resulted in further improvements according to both AIC and BIC. In fact, the more complex models even show slight decreases in F1 scores. Thus, among the linguistic features considered, only subcategorization probability enhanced the predictions of CC presence.

After applying Lasso Regression, four of seven features were retained: subcategorization probability, verb frequency, factivity, and subject form. Using this refined set, we trained a structural predictability model with the same neural network architecture. However, although this model achieved a lower log loss (0.4790), the model showed a decrease in F1 score (0.6519) and an increase in BIC compared to the model using only Subcategorization Probability. This result is consistent with earlier findings from incremental feature selection, further confirming that additional linguistic features do not improve predictive performance.

In contrast, the model trained on contextual word embedding features achieved a log loss of 0.442 and an F1 score of 0.6906, outperforming all models based on hand-selected features.

Based these results, we proceeded to test how well information density derived from (i) verb subcategorization probabilities and (ii) from contextual word embeddings predicts *that*-mentioning.

4 Information Density and *that*-Mentioning

This section reports our statistical models predicting *that*-mentioning. We examined whether higher information density—estimated from verb subcategorization probabilities and contextual word embeddings—leads to increased *that*-mentioning, as predicted by UID. Additionally, we assessed whether more accurate estimates of CC structural predictability improve the overall fit of models predicting *that*-mentioning.

Features	AIC Δ	BIC Δ	F1	Log Loss
Intercept only	–	–	0.000	0.6346
Subcategorization Probability	-13629.10	-13629.10	0.6598	0.4905
+ Verb Frequency	128.94	1456.78	0.6598	0.4892
+ Factivity	320.06	1647.90	0.6598	0.4912
+ Tense	156.09	1483.93	0.6541	0.4895
+ Position	247.66	1575.50	0.6596	0.4904
+ Subject Form	-313.63	1014.20	0.6470	0.4845
+ Subject Frequency	-514.77	813.07	0.6451	0.4824

Table 1: Model comparisons predicting CC presence. Lower AIC, BIC, and log loss, and higher F1 scores indicate better performance.

4.1 Data

As in previous analyses, we relied on parsed transcripts from the CANDOR corpus (Reece et al., 2023). We extracted CC introduced by the same 50 matrix verbs used for training CC structural predictability models (Appendix A) and retained only instances where the matrix verb preceded the CC.

The dataset was further refined based on the following criteria. First, we excluded the first CCs in all conversations (1,656 cases), as we are interested in the potential effects of whether the previous CC is reduced or not. Second, we removed cases lacking either a matrix subject or an embedded nominal subject, as the identity of both the matrix and the embedded subjects are crucial for our analysis (13,076 cases). Lastly, for matrix verbs introducing multiple CCs, only the first occurrence was retained to avoid redundancy (8,097 cases excluded). After exclusions, we are left with 51,276 instances of CCs for analysis.

4.2 Control variables

To rigorously test UID predictions, we controlled for a range of variables that can also affect *that*-mentioning, largely following Jaeger (2010). We discuss each of them in the following subsections. Importantly, the UID account is not mutually exclusive with these mechanisms. See Appendix B for a summary of the control variables, including their types, levels, and relative proportions.

4.2.1 Availability-Based Production

According to availability-based accounts (Bock and Warren, 1985; Ferreira, 1996; Ferreira and Dell, 2000), optional elements facilitate production when upcoming material is less accessible (i.e., when upcoming material has low frequency). To capture such effects, we included the log frequency of the CC subject head (CC SUBJECT FREQUENCY), the form of the CC subject (CC SUBJECT FORM; *I* vs. *You* vs. *Other pronouns* vs. *Other nouns*), and the matrix verb's log frequency (MATRIX VERB FREQUENCY). We also included CO-REFERENTIALITY, a binary predictor indicating whether the matrix and CC subjects are identical (e.g., *I think I...*).

4.2.2 Syntactic Priming

Speakers tend to repeat recently encountered structures (Bock, 1986; Gries, 2005; Mahowald et al., 2016). We included PREVIOUS THAT, a binary predictor indicating whether *that* was present in the speaker's or interlocutor's most recent CC.

4.2.3 Dependency Locality

Longer dependencies increase production difficulty (Hawkins, 2004; Roland et al., 2006). Three locality measures were considered: MATRIX VERB-CC DISTANCE (local vs. non-local), CC SUBJECT LENGTH (number of the CC subject's dependents), and CC REMAINDER LENGTH (number of words following the CC subject head in the same CC).

4.2.4 Speaker Commitment

It has been argued that variation in *that*-mentioning is not meaning-equivalent (Thompson and Mulac, 1991), as sometimes the matrix verb conveys the speaker's level of commitment rather than introducing a true CC, making *that* unnecessary. Following Jaeger (2010), we assumed that commitment is highest with first-person subjects, followed by second-person, and then third-person references, and included MATRIX SUBJECT FORM as a four-level predictor (*I* vs. *You* vs. *Other pronouns* vs. *Other nouns*).

4.2.5 Position

Effects Production difficulty may vary depending on when the CC occur in a sentence. We included

VERB ID, the ordinal position of the matrix verb, as a continuous predictor.

4.2.6 Similarity Avoidance

Speakers may omit *that* to avoid adjacent similar forms if the CC also begins with *that* (Walter and Jaeger, 2008). We included THAT-DOUBLING as a binary predictor.

4.2.7 Disfluencies

Disfluencies can impact syntactic choices (e.g., Liang et al., 2024). We included FILLED WORD (presence of a filled pause before the CC) and REPETITION (immediate repetition of a word, excluding adjectives and adverbs used for emphasis).

4.3 Statistical Modeling of *that*-mentioning

Before modeling, all continuous predictors (see Appendix B) were standardized using z-score normalization. Binary predictors were contrast-coded, and the four-level categorical variables (CC SUBJECT FORM and MATRIX SUBJECT FORM) were coded using successive difference coding: comparing *I* vs. *You*, *You* vs. *Other Pronouns*, and *Other Pronouns* vs. *Other Nouns*.

We fitted a generalized linear mixed-effects model (GLMM; Jaeger, 2008) using the glmer() function from the lme4 package in R (Bates et al., 2015; R Core Team, 2023), with the presence of *that* as the binary dependent variable. Fixed effects included CC information density and a set of control variables. To account for variability across individuals, we included a random intercept for speaker. In follow-up analyses, we also included a random intercept for matrix verb lemmas (verbs henceforce) to capture verb-specific tendencies in complementizer usage. While these random effects are not directly motivated by theoretical accounts, they serve to control for idiosyncratic variation in baseline rates of *that*-mentioning across speakers and lexical items. Model comparisons were evaluated via AIC and BIC.

4.4 Results of *that*-Mentioning

Here we report the effects of information density on *that*-mentioning to test predictions from the UID hypothesis, alongside other control variables. Information density was estimated using two approaches: the matrix verb's subcategorization probability and its contextual word embedding. We further examine whether embedding-based estimates—shown to more accurately predict CC presence—better account for *that*-mentioning patterns than verb-based estimates.

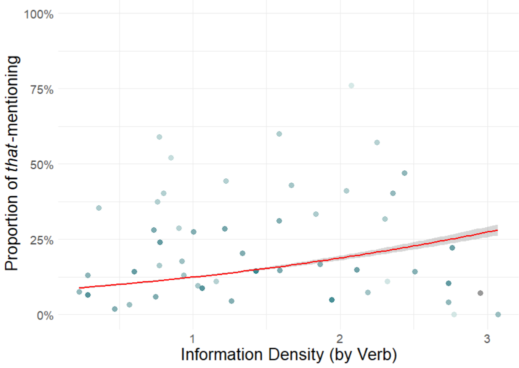

Figure 2: Effects of information density (by verb Subcategorization Probability) on *that*-mentioning.

4.4.1 Verb-based Information Density

The relationship between verb-based information density and *that*-mentioning is illustrated in Figure 2. Higher information density is generally associated with increased rates of *that*-mentioning, consistent with UID predictions, although substantial variability across verbs remains. Results from the statistical model with a speaker random intercept are presented in Table 2. Generalized Variance Inflation Factors (GVIFs) for fixed effects were close to 1, indicating minimal multicollinearity. Mode results revealed that higher verb-based information density significantly increases the likelihood of *that*-mentioning.

Effects of control variables also aligned with several theoretical accounts. First, higher CC SUBJECT FREQUENCY and MATRIX VERB FREQUENCY predicted reduced *that*-mentioning, consistent with availability-based accounts. However, CC SUBJECT FORM and CO-REFERENTIALITY were non-significant. We also found syntactic priming effects, whereby PREVIOUS *that* significantly increased *that*-mentioning. Findings for dependency locality were mixed: longer CC REMINDER LENGTH increased *that*-mentioning as expected, but shorter CC SUBJECT LENGTH and MATRIX VERB-CC DISTANCE also led to higher *that*-use, contrary to the predictions. Speaker commitement effects were robust—*that* was more likely when the matrix subject was *You* than *I*, with similar trends across other subject types, suggesting *that* signals degrees of speaker commitment. VERB ID had a positive but non-significant effect. Supporting similarity avoidance, potential *that*-DOUBLING reduced *that*-mentioning. Finally, disfluencies measures

Predictor	Estimate	p-value
Information Density	0.28	< 0.001
CC Subject Frequency	-0.16	< 0.001
CC Subject Form 2–1	-0.00	$= 0.99$
CC Subject Form 3–2	-0.02	$= 0.72$
CC Subject Form 4–3	0.03	$= 0.68$
Matrix Verb Frequency	-0.24	< 0.001
Co-referentiality	-0.05	$= 0.24$
Previous *that*	0.23	< 0.001
Matrix Verb-CC Distance	-0.40	< 0.001
CC Subject Length	-0.04	< 0.05
CC Reminder Length	0.18	< 0.001
Matrix Subject Form 2–1	0.69	< 0.001
Matrix Subject Form 3–2	0.70	< 0.001
Matrix Subject Form 4–3	0.53	< 0.001
Verb ID	0.02	$= 0.20$
that-Doubling	-0.53	< 0.001
Filled Word	0.04	$= 0.36$
Repetition	0.14	< 0.05

Table 2: Regression estimates from the model predicting complementizer presence.

such as FILLED WORD and REPETITION increased *that*-use, with REPETITION reaching significance.

4.4.2 Embedding-based Information Density

As shown in Figure 3, embedding-based information density again positively predicted *that*-mentioning. Because the statistical results closely mirrored those of the previous model, we do not report them in detail. Crucially, information density remained a strong predictor ($\beta = 0.15$; $p < 0.001$).

However, the current model performed worse, with AIC and BIC increasing by 392 and 391 points, respectively, compared to the previous model with verb-based information density. While word embedding features yielded better performance in the structural predictability task, they offered no clear advantage in predicting *that*-mentioning over subcategorization probabilities.

4.5 Follow-up Analysis: Verb Random Intercept

Although the verb-based model initially outperformed the embedding-based model, we were cautious in interpreting this as evidence that embedding-based information density is less effective. In the verb-based model, information density is constant for each matrix verb, potentially conflating information density with verb-specific effects—a limitation of subcategorization probability

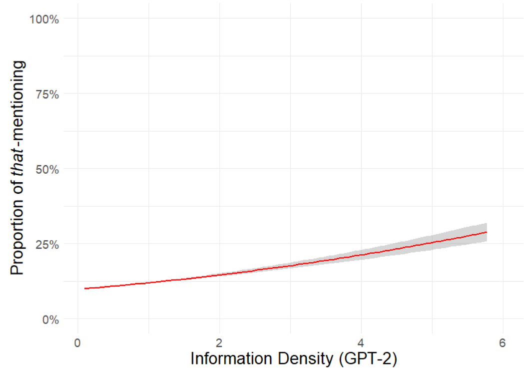

Figure 3: Effects of information density (by contextual word embeddings) on *that*-mentioning.

we mentioned earlier. To address this, we refitted both models with an added random intercept for matrix verbs.

We found that adding a matrix verb random intercept substantially reduced AIC and BIC for both the verb-based and embedding-based models (Table 3), indicating that a substantial portion of variation in complementizer usage is attributable to verb-specific preferences—patterns tied to individual matrix verbs that were not captured by fixed effects in the previous models.

Additionally, the effects of information density diverged. In the verb-based model, the effect of information density became non-significant ($\beta = 0.14$; $p = 0.18$), suggesting that its earlier effect was largely driven by verb-specific variation. In contrast, information density estimated from contextual word embeddings remained a significant predictor even after controlling for verb identity ($\beta = 0.12$; $p < 0.001$). Furthermore, between the two models with verb random intercepts, the embedding-based model showed better fit, reducing AIC and BIC by 25 and 26 points, respectively, suggesting that embedding-based information density captures additional variance in patterns of *that*-mentioning.

5 Discussion

This study revisited Jaeger (2010) using a large and modern dataset from the CANDOR corpus. We analyzed over 50,000 instances of CCs to test how information density—estimated from different sources—predicts *that*-mentioning, alongside other predictors motivated by alternative theories. We also evaluated whether improved estimates of information density lead to better model performance.

Our results replicated the core finding that higher

Model	AIC	BIC
Verb-based information density, without verb random intercept	33344	33521
Embedding-based information density, without verb random intercept	33736	33912
Verb-based information density, with verb random intercept	32302	32488
Embedding-based information density, with verb random intercept	**32277**	**32462**

Table 3: Model comparison based on AIC and BIC.

information density increases the likelihood of overt *that*, as predicted by UID. Information density estimated from verb subcategorization probabilities provided strong predictive power but likely reflected verb-specific preferences rather than a general effect of information density. This was confirmed by follow-up models with random intercepts for matrix verbs, which eliminated the effect of verb-based information density. In contrast, embedding-based information density remained significant in predicting *that*-mentioning, suggesting it captures more abstract, verb-independent information. Moreover, this is consistent with the results of structural predictability models, where GPT-2 embeddings did outperform all other features, including verb subcategorization probability, in predicting CC presence, suggesting that it offers a better measure of information density.

However, we do note that after including the verb random intercept, Jaeger (2010) still found significant effects of verb-based measures of information content. This discrepancy may be attributed to differences in dataset size and verb diversity. Jaeger's (2010) study was based on on a smaller dataset with a more limited set of verbs, which may have amplified the observed effects

Beyond UID, we also found support for other accounts of *that*-mentioning. First, lower-frequency matrix verbs and CC onsets were associated with more *that*-mentioning, consistent with availability-based accounts. We also found syntactic priming: speakers were more likely to include *that* if the previous CC did. Evidence for dependency locality was mixed—longer CC remainders increased *that*-mentioning, but greater distance between the matrix verb and CC onset, as well as longer CC subjects, showed the opposite pattern. This may be due to parsing errors or shifting usage patterns. Effects of speaker commitement were also observed, with higher levels of speaker commitment leading to less overt *that*. Finally, we observed similarity avoidance (reduced *that*-use in potential *that-that* sequences) and disfluency effects (filled words and repetitions increased *that*-mentioning).

Our findings also shed light on the structural sensitivity of GPT-2, particularly its contextual word embeddings. Embeddings of the matrix verb—derived solely from pre-CC context—were predictive of upcoming syntactic structure, suggesting that GPT-2 captures fine-grained structural cues. This approach offers a promising avenue for future work to leverage contextual embeddings for modeling syntactic prediction more broadly.

6 Conclusion

This study provides robust support for UID at the syntactic level in naturalistic conversations. Information density estimated from contextual word embeddings significantly predicted *that*-mentioning, even after controlling for verb-specific preferences. Additionally, we showed that verb-specific preferences also played an important role, and that information density measures derived from verbs' subcategorization probabilities might have been conflated with verb-specific preferences. These findings highlight limitations of conventional linguistic features in modeling predictive processes, and suggest that high-dimensional linguistic representations such as contextual word embeddings offer a more effective and flexible alternative. Our results also demonstrate that *that*-reduction is shaped by multiple interacting pressures—including information density, availability, speaker commitment, syntactic priming, and form avoidance.

Lastly, our work underscores the value of combining large naturalistic corpora with machine learning and NLP techniques for studying psycholinguistics. The use of the CANDOR corpus allowed us to examine *that*-mentioning in spontaneous, naturalistic speech across a diverse linguistic samples. By leveraging machine learning and contextual word embeddings from neural language models, we developed more nuanced predictors of structural choices. This approach not only improves predictive accuracy but also opens new avenues for modeling linguistic behavior at scale.

Limitations

There are several limitations to the present study. First, the conversational transcripts were automatically generated, and dependency structures were derived using automatic parsers. As a result, the data may contain transcription and parsing errors. Second, we relied on GPT-2 to estimate online spoken language predictions, although GPT-2 is primarily trained on written text. This may limit its ability to fully capture characteristics of spontaneous spoken language. Moreover, our analysis was based on a single language model architecture. Future work should explore alternative models, including those trained on conversational data or designed for speech-oriented tasks, to assess the generalizability of our findings. Lastly, although our analysis found that no linguistic features significantly improved the structural predictability of complementizer clauses, it is possible that we did not exhaust the full range of relevant linguistic predictors. Future research could investigate additional features that may contribute to CC presence.

Ethical Considerations

We employed AI-based tools (Claude and ChatGPT) for writing and coding assistance. These tools were used in compliance with the ACL Policy on the Use of AI Writing Assistance.

References

Hirotugu Akaike. 1974. A new look at the statistical model identification. In *Springer Series in Statistics*, pages 215–222. Springer.

Suhas Arehalli and Tal Linzen. 2020. Neural language models capture some, but not all, agreement attraction effects. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 370–376. Cognitive Science Society.

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform information density at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Douglas Bates, Martin Mächler, Ben Bolker, and Steven Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Alan Bell, Jason M. Brenier, Michelle Gregory, Cynthia Girand, and Daniel Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111.

J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Kathryn Bock and Rose Warren. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21(1):47–67.

Marc Brysbaert and Boris New. 2009. Moving beyond kucera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, 41(4):977–990.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT‘s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Uriel Cohen Priva. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, 6(2).

Marie-Catherine de Marneffe, Joakim Nivre, Mitchell Abrams, Cristina Bosco, Richárd Farkas, Hiroshi Kanayama, Seungyoung Kang, Natalia Kotsyba, Veronika Laippala, Teresa Lynn, Christopher D. Manning, Akshat Minocha, Anna Missilä, Stephan Oepen, Sameer Pradhan, Sampo Pyysalo, Natalia Silveira, Katarzyna Szał, Reut Tsarfaty, and Daniel Zeman. 2021. Universal Dependencies v2: An evergrowing multilingual treebank collection. *Computational Linguistics*, 47(2):255–308.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2024. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Victor S. Ferreira. 1996. Is it better to give than to donate? syntactic flexibility in language production. *Journal of Memory and Language*, 35(5):724–755.

Victor S. Ferreira and Gary S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40(4):296–340.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*, pages 517–520.

Stefan Gries. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, 34(4):365–399.

Hailin Hao. 2023. Evaluating transformers' sensitivity to syntactic embedding depth. In Rashid Mehmood and 1 others, editors, *Distributed Computing and Artificial Intelligence, Special Sessions I, 20th International Conference (DCAI 2023)*, volume 741 of *Lecture Notes in Networks and Systems*, pages 175–182. Springer, Cham.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. `https://spacy.io`. Software available from https://spacy.io.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

T. Florian Jaeger. 2008. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

T. Florian Jaeger and Jane Grimshaw. 2013. Information density affects both production and grammatical constraints. In *Proceedings of the Architectures and Mechanisms for Language Processing (AMLaP) Conference*, Marseille, France.

Ian T. Jolliffe. 2002. *Principal Component Analysis*, 2nd edition. Springer.

Lauri Karttunen. 1971. Implicative verbs. *Language*, 47(2):340–358.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition: A reply to piantadosi (2023). Manuscript, Tel Aviv University. Available at LingBuzz: `https://lingbuzz.net/lingbuzz/007190`.

Bobak Kennedy, Anjali Ashokkumar, Ryan L. Boyd, and Morteza Dehghani. 2021. Text analysis for psychology: Methods, principles, and practices. In Morteza Dehghani and Ryan L. Boyd, editors, *Handbook of Language Analysis in Psychology*, pages 3–62. The Guilford Press.

Victor Kuperman, Mark Pluymaekers, Mirjam Ernestus, and R. Harald Baayen. 2007. Morphological predictability and acoustic duration of interfixes in dutch compounds. *Journal of the Acoustical Society of America*, 121(4):2261–2271.

Ramon Lemke, Emina Horch, and Ingo Reich. 2017. Optimal encoding! – information theory constrains article omission in newspaper headlines. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 131–135. Association for Computational Linguistics.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 19. MIT Press.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Yuhan Liang, Pascal Amsili, Heather Burnett, and Vera Demberg. 2024. Uniform information density explains subject doubling in french. In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society (CogSci)*.

Tomasz Limisiewicz and David Mareček. 2020. Syntax representation in word embeddings and neural networks: A survey. In *Proceedings of the 20th Conference ITAT 2020: Automata, Formal and Natural Languages Workshop*.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Kyle Mahowald, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Kyle Mahowald, Adam James, Richard Futrell, and Edward Gibson. 2016. A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, 91:5–27.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3. LDC99T42.

James Michaelov and Benjamin Bergen. 2020. How well does surprisal explain n400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Jackson Petty, Michael Wilson, and Robert Frank. 2022. Do language models learn position-role mappings? In *Proceedings of the 46th Annual Boston University Conference on Language Development (BUCLD)*, volume 2, pages 657–671, Somerville, MA. Cascadilla Press.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mark Pluymaekers, Mirjam Ernestus, and R. Harald Baayen. 2005. Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America*, 118(4):2561–2569.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Technical report, OpenAI.

Andrew Reece, Gabe Cooney, Peter Bull, Carolyn Chung, Benjamin Dawson, Charles Fitzpatrick, Talia Glazer, Daniel Knox, Alexandra Liebscher, and Sophie Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13).

Douglas Roland, Jeffrey L. Elman, and Victor S. Ferreira. 2006. Why is that? structural prediction and ambiguity resolution in a very large corpus of english sentences. *Cognition*, 98(3):245–272.

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Sandra A. Thompson and Anthony Mulac. 1991. A quantitative perspective on the grammaticalization of epistemic parentheticals in english. In *Typological Studies in Language*, volume 19, pages 313–339. John Benjamins.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Martha A. Walter and T. Florian Jaeger. 2008. Constraints on optional *that*: A strong word form ocp effect. In *Proceedings of the Main Session of the 41st Meeting of the Chicago Linguistic Society*, pages 505–519, Chicago, IL. Chicago Linguistic Society.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Michael Wilson, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

Meilin Zhan and Roger Levy. 2018. Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1997–2005, New Orleans, Louisiana. Association for Computational Linguistics.

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. How well do large language models understand syntax? an evaluation by asking natural language questions. *Preprint*, arXiv:2311.08287.

A Matrix Verb Statistics

This appendix provides the full distribution of complement clause subcategorization probabilities across verbs in our dataset in Table 4.

B Descriptive Statistics for Predictors for Modeling *that*-Mentioning

This appendix provides the full distribution of complement clause subcategorization probabilities across verbs in our dataset in Table 5.

Verb Lemma	Total Occurrences	CC Occurrences	Subcat Probability (%)
know	119,678	28,664	23.95
think	46,610	35,080	75.26
mean	30,281	1,916	6.33
say	24,805	13,612	54.88
like	23,578	3,381	14.34
see	20,578	7,111	34.56
take	15,314	994	6.49
feel	11,274	2,298	20.38
guess	9,744	6,101	62.61
hear	9,166	2,408	26.27
tell	7,264	3,345	46.05
find	6,579	1,948	29.61
love	6,290	762	12.11
thank	5,521	289	5.23
remember	4,626	2,191	47.36
read	3,649	346	9.48
show	3,170	650	20.50
understand	2,984	1,092	36.60
suppose	2,911	326	11.20
hope	2,488	1,869	75.12
teach	2,380	238	10.00
figure	2,327	912	39.19
believe	1,970	947	48.07
imagine	1,891	874	46.22
check	1,754	114	6.50
care	1,693	263	15.53
decide	1,428	579	40.55
realize	1,395	974	69.82
agree	1,324	172	12.99
hold	1,313	107	8.15
wish	1,291	1,028	79.63
worry	1,028	90	8.75
expect	980	349	35.61
consider	840	264	31.43
mind	733	208	28.38
notice	721	324	44.94
mention	645	190	29.46
answer	561	26	4.63
explain	561	106	18.89
bet	480	272	56.67
accept	465	49	10.54
complain	423	53	12.53
stress	234	23	9.83
admit	209	98	46.89
respond	176	11	6.25
joke	156	32	20.51
promise	146	58	39.73
judge	119	19	15.97
claim	110	47	42.73
suggest	108	50	46.30

Table 4: Verb-level complement clause frequencies and subcategorization probabilities.

Predictor	Type	Values / Distribution
CC Subject Frequency	Continuous	–
CC Subject Form	Categorical (4 levels)	I (29.62%), You (13.15%), other pronouns (41.80%), other NPs (15.42%)
Matrix Verb Frequency	Continuous	–
Co-referentiality	Binary	yes (32.72%), no (67.28%)
Previous *that*	Binary	present (11.46%), absent (88.54%)
Matrix Verb-CC Distance	Binary	local (84.73%), non-local (16.27%)
CC Subject Length	Continuous	–
CC Reminder Length	Continuous	–
Matrix Subject Form	Categorical (4 levels)	I (72.70%), You (10.37%), other pronouns (13.49%), other NPs (3.44%)
Position	Continuous	–
that-Doubling	Binary	present (3.03%), absent (96.97%)
Filled Word	Binary	present (10.32%), absent (89.68%)
Repetition	Binary	present (4.12%), absent (95.88%)

Table 5: Overview of predictors included in the statistical model, along with their types and distribution where applicable.