Peak Information Density and Complementizer Realization in Learner English: Extending Uniform Information Density Through Local Information-Density Peaks

Yuichi Ono, Reina Mogushi, Shota Kato, Yuka Fujii, and Kasumi Takahashi
Institute of Humanities and Social Sciences , Degree Programs in Humanities and Social Sciences
University of Tsukuba
ono.yuichi.ga@u.tsukuba.ac.jp

Abstract

Optional grammatical markers provide an important testing ground for investigating how speakers manage information during language production. Previous psycholinguistic and corpus-based studies have shown that English complementizer realization is influenced by processing difficulty and information density (Ferreira & Dell, 2000; Jaeger, 2005, 2010). However, it remains unclear which temporal and distributional profile of information density is most relevant to optional complementizer realization. Existing studies have primarily focused on clause-initial predictability or overall information density, while the role of localized informational bottlenecks has received little attention. The present study addresses this issue using 4,954 complement clauses extracted from the written and spoken components of the International Corpus Network of Asian Learners of English (ICNALE). Word-level surprisal values were estimated using GPT-2 to derive competing information-density measures, including clause-initial, cumulative, mean, total, and peak surprisal. Generalized linear mixed-effects models were used to compare their ability to predict overt that realization. The results showed that peak surprisal significantly predicted complementizer realization independently of clause-initial and mean surprisal and provided the best overall model fit. Furthermore, the effect was stronger in written production than in spoken production. A supplementary analysis of spoken data found no evidence that the effect was mediated by longer pauses or complementizer duration, suggesting that overt that primarily functions as a structural signal rather than a temporal buffering device. These findings demonstrate the value of corpus-based surprisal modeling for identifying the informational profile underlying optional grammatical choice and support the Peak Information Density Hypothesis (PIDH) as a refinement of the Uniform Information Density framework.

Keywords: Uniform Information Density, Peak Information Density Hypothesis, Complementizer Realization, GPT-2 Surprisal, Learner Corpus

1. Introduction

Many grammatical phenomena allow speakers to choose between alternative forms without substantially altering propositional meaning. Such optional grammatical variation provides a valuable window into the cognitive mechanisms underlying language production because speakers must continuously balance production efficiency with communicative effectiveness. One of the best-known examples is the English complementizer that, which may be either overtly realized or omitted following predicates such as *think*, *believe*, and *say* (e.g., *I think (that) he is right*). Because both variants express essentially the same proposition, complementizer realization has long been regarded as an ideal testing ground for investigating the factors governing probabilistic grammatical choice.

Previous research has demonstrated that complementizer realization is influenced by a wide range of linguistic and cognitive factors. Structural accounts have emphasized syntactic complexity, ambiguity avoidance, and dependency length (Hawkins, 2004), whereas psycholinguistic studies have argued that overt *that* reflects increased production difficulty or planning demands associated with the upcoming complement clause (Ferreira & Dell, 2000; Jaeger, 2005). Under incremental models of language production (Levelt, 1989), speakers are assumed to plan upcoming material while producing the current utterance, making optional complementizers a useful indicator of production planning and processing difficulty. These studies collectively suggest that optional grammatical markers emerge not randomly but as adaptive responses to the cognitive demands imposed by upcoming linguistic material.

A broader explanation has been provided by the Uniform Information Density (UID) hypothesis (Levy & Jaeger, 2007; Jaeger, 2010), which proposes that speakers distribute information relatively evenly across an utterance in order to maximize communicative efficiency while avoiding excessive local processing difficulty. Within this framework, optional grammatical material serves to regulate information flow when highly informative material is expected. Previous studies have therefore interpreted overt *that* as one mechanism for smoothing information density before the onset of an upcoming complement clause.

Despite these advances, an important theoretical question remains unresolved. Although previous studies agree that information density influences complementizer realization, it remains unclear which aspect of information density is actually relevant to speakers' grammatical choices. Information density can be quantified in several ways, including the predictability of the clause-initial word, the cumulative informational load during clause onset, the average information density across the clause, or localized peaks of surprisal occurring within the clause. Existing studies have rarely compared these alternative informational profiles directly, leaving the temporal locus of information management largely unspecified.

Addressing this question requires evidence from naturally occurring language. Controlled psycholinguistic experiments have been indispensable for identifying causal relationships between production difficulty and

grammatical choice, but experimental paradigms necessarily manipulate a limited set of predetermined variables. In contrast, corpus linguistics makes it possible to examine thousands of naturally produced clauses and compare multiple competing measures of information density within the same linguistic environment. Recent advances in neural language models further enable word-level surprisal to be estimated directly from authentic language production, providing an effective means of evaluating competing information-density hypotheses in large-scale learner corpora.

The present study adopts this corpus-based approach using the written and spoken components of the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2023). Word-level surprisal values estimated by GPT-2 were used to derive five competing measures of information density: clause-initial, cumulative, mean, total, and peak surprisal. By comparing these measures within a unified generalized linear mixed-effects modeling framework, the study investigates which informational profile best predicts optional complementizer realization. Based on the findings, we propose the Peak Information Density Hypothesis (PIDH), which argues that complementizer realization is primarily associated with localized peaks of information density within the upcoming complement clause.

2. Related Work

2.1 Optional Complementizer Realization

The English complementizer *that* is one of the most extensively investigated cases of optional grammatical variation because its realization is generally not required for successful communication. Speakers may either produce or omit the complementizer without substantially changing propositional meaning (e.g., *I think (that) he is right*). This optionality has made complementizer realization an important testing ground for theories of probabilistic language production.

Previous studies have identified numerous factors influencing complementizer realization. Structural accounts have emphasized dependency length, syntactic complexity, and ambiguity avoidance (Hawkins, 2004), whereas psycholinguistic approaches have argued that overt *that* reflects increased production difficulty associated with planning the upcoming complement clause (Ferreira & Dell, 2000). Jaeger (2005, 2010) further demonstrated that complementizer realization is systematically related to speakers' expectations about upcoming linguistic material, suggesting that optional grammatical markers function as adaptive responses to processing demands rather than arbitrary stylistic variation.

In summary, these studies indicate that optional complementizer realization is sensitive to characteristics of the upcoming clause. However, they do not specify which representation of upcoming processing difficulty best predicts speakers' probabilistic choice.

2.2 Uniform Information Density and Surprisal

The Uniform Information Density (UID) hypothesis (Levy & Jaeger, 2007; Jaeger, 2010) provides a general theoretical account of optional grammatical variation. According to UID, speakers tend to distribute information relatively evenly across an utterance in order to maximize communicative efficiency while minimizing local processing difficulty. Under this view, optional grammatical markers such as the complementizer *that* are expected to emerge when they help regulate information flow before cognitively demanding linguistic material.

To evaluate this theoretical claim, however, information density must be operationalized in a measurable form. Surprisal, defined as the negative logarithm of the conditional probability of a word given its preceding context (Hale, 2001; Levy, 2008), has become the standard cognitive metric for quantifying information density. Higher surprisal values indicate lower predictability and greater processing difficulty, making surprisal an empirically tractable indicator of the cognitive demands assumed by UID.

Recent advances in neural language models have substantially improved the estimation of surprisal. Unlike earlier *n*-gram models, transformer-based language models capture long-range contextual dependencies and provide more accurate estimates of word predictability from naturally occurring language. Consequently, models such as GPT-2 enable researchers to estimate word-level surprisal directly from large-scale corpora and have become increasingly common in psycholinguistic and corpus-linguistic research.

It is important to note that these three components serve different roles in the present study. UID provides the theoretical framework, surprisal serves as the cognitive measure of information density, and GPT-2 functions as the computational method for estimating surprisal from corpus data. Distinguishing these levels is essential because the theoretical predictions concern information density itself rather than any particular language model. GPT-2 is employed not as a theory of language production, but as a practical tool for approximating the probabilistic expectations assumed by UID.

However, while surprisal has become the standard operationalization of information density, it remains unclear which representation of surprisal is most relevant to optional grammatical choice. Previous studies have typically examined clause-initial predictability or average information density, whereas alternative representations—including cumulative, total, and localized peak surprisal—have rarely been compared systematically. The present study addresses this gap by evaluating multiple competing surprisal measures within a unified corpus-based framework.

2.3 Corpus-Based Investigation of Information Density

Experimental psycholinguistic studies have played a central role in demonstrating that optional grammatical choices are influenced by production difficulty and linguistic expectation. Their principal advantage lies in the ability to isolate specific causal factors under carefully controlled conditions. However, experimental paradigms necessarily manipulate a limited number of predetermined variables and therefore cannot fully capture the rich variability found in naturally occurring language.

Corpus linguistics provides a complementary perspective by enabling researchers to examine large numbers of

naturally produced utterances in which lexical, syntactic, and discourse-related factors interact simultaneously. Combined with neural language models, corpus-based surprisal analysis allows multiple competing representations of information density to be evaluated within the same dataset. Rather than asking whether information density influences optional grammatical realization, corpus-based approaches make it possible to investigate which informational profile best accounts for naturally occurring grammatical choice.

The present study adopts this corpus-based perspective by comparing five competing measures of information density derived from GPT-2 surprisal estimates. In doing so, it aims to identify the informational profile underlying optional complementizer realization and to evaluate the Peak Information Density Hypothesis (PIDH).

3. Method

3.1 Corpus Revision Pipeline

The study used the written component of the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2023). The analysis focused on finite complement clauses following matrix predicates that optionally permit the realization of the complementizer *that*. Each complement clause was automatically extracted together with information on writer identity and sentence ID, allowing each observation to be linked to both linguistic and speaker-level metadata.

The dependent variable was the presence or absence of the overt complementizer *that*.

3.2 Estimating Surprisal with GPT-2

To operationalize information density, word-level surprisal was estimated using the pretrained GPT-2 language model (Radford et al., 2019) implemented through the Hugging Face Transformers library.

Following Hale (2001) and Levy (2008), surprisal was defined as the negative logarithm of the conditional probability of each word given its preceding context:

$$\mathbf{Surprisal}(w_i) = -\log P\,(w_i|w_1, \dots, w_{i-1})$$

For each complement clause, GPT-2 generated token-level probability distributions over the vocabulary. The log probability assigned to the observed next token was extracted, and its negative log value was taken as the surprisal of that token. This procedure yielded a sequence of token-level surprisal values for each complement clause.

3.3 Information-Density Measures

To determine which aspect of information density best predicts complementizer realization, five competing summary measures were derived from the token-level surprisal sequence.

For each complement clause, token-level surprisal values were represented as

$$S = (s_1, s_2, \dots, s_n).$$

where *n* denotes the number of tokens in the complement clause. Five summary measures were computed from this surprisal sequence.

- **Clause-initial surprisal**

The surprisal contained in the beginning of the clause was summarized using cumulative surprisal over the first one, three, and five tokens:

$$\mathbf{Surp1} = s_1$$

$$\mathbf{Surp3} = s_1 + s_2 + s_3$$

$$\mathbf{Surp5} = s_1 + s_2 + s_3 + s_4 + s_5$$

These measures represent increasingly broader estimates of the informational load encountered immediately after complement-clause onset.

- **Mean surprisal**

The average information density across the clause was calculated as

$$\mathbf{Mean} = (s_1 + s_2 + \cdots + s_n)/n$$

This measure corresponds to the overall information density assumed in many UID-based analyses.

- **Total surprisal**

The total informational load of the clause was calculated as

$$\mathbf{Total} = s_1 + s_2 + \cdots + s_n$$

Unlike the mean, this measure reflects the cumulative processing demand across the entire complement clause.

- **Peak surprisal**

To capture localized informational bottlenecks, peak surprisal was defined as

$$\mathbf{Peak} = \max(s_1, s_2, \dots, s_n)$$

Rather than representing average processing difficulty, peak surprisal identifies the single most unexpected token within the complement clause.

3.4 Statistical Analysis

Each information-density measure was entered separately into generalized linear mixed-effects models predicting complementizer realization.

The dependent variable was binary:

- 1 = overt *that*
- 0 = zero complementizer

Writer was included as a random intercept to account for repeated observations from the same individual. Model performance was compared across competing surprisal measures using Akaike's Information Criterion (AIC), with lower AIC values indicating better model fit.

The primary objective was not merely to determine whether information density predicts complementizer realization, but rather which representation of information density provides the best account of naturally occurring optional grammatical choice.

In addition to fitting separate models for each information-density measure, combined models were fitted with Surp1, Mean surprisal, and Peak surprisal entered simultaneously. A further model included the interaction between Peak surprisal and production mode. All models controlled for L1 background, proficiency level, and production mode, and included a random intercept for writer.

4. Results

4.1 Comparison of Information-Density Measures

To determine which representation of information density best predicted overt complementizer realization, a series of generalized linear mixed-effects models was compared using Akaike's Information Criterion (AIC). All models included production mode, L1 background, and proficiency level as control variables, together with a random intercept for writer.

The final dataset contained 4,954 complement-clause observations produced by 397 learners. Overt that occurred in 1,341 observations, corresponding to an overall realization rate of 27.1%. The rate was substantially higher in the written data than in the spoken data: 35.5% in writing and 13.6% in speech.

Table 1 presents the model comparison. Among the single-measure models, the model containing Peak surprisal provided the best fit, with an AIC of 5127.02. It slightly outperformed the clause-initial Surp1 model, whose AIC was 5132.54, and clearly outperformed the Total and Mean surprisal models.

However, the best overall fit was obtained from the combined model including Surp1, Mean surprisal, Peak surprisal, and the interaction between Peak surprisal and production mode. This model yielded an AIC of 4964.42, which was 162.61 points lower than the Peak-only model. The additive model containing Surp1, Mean, and Peak surprisal also performed substantially better than any single-measure model.

Model	Information-density predictors	df	AIC	ΔAIC
Combined × Mode	Surp1 + Mean + Peak × Mode	12	**4964.42**	**0**
Combined	Surp1 + Mean + Peak	11	4972.05	7.63
Peak	Peak	9	5127.02	162.61
Initial	Surp1	9	5132.54	168.12
Total	Total	9	5180.27	215.85
Mean	Mean	9	5199.78	235.36

Table 1: Comparison of GLMMs predicting overt complementizer realization

A likelihood-ratio test confirmed that the combined model provided a significantly better fit than the Peak-only model, $\chi^2(2) = 158.97$, $p < .001$. Adding the interaction between Peak surprisal and production mode further improved model fit, $\chi2(1) = 9.63$, $p = .002$.

These results indicate that Peak surprisal was the strongest individual information-density measure, but that complementizer realization was best explained by a combination of clause-initial, mean, and peak information density.

4.2 Effects of Initial, Mean, and Peak Surprisal

Table 2 presents the fixed effects from the best-fitting model.

Clause-initial surprisal significantly increased the likelihood of overt complementizer realization, b = 0.331, SE = 0.038, $z = 8.63$, $p < .001$. Because the surprisal predictors were standardized, this coefficient indicates that a one-standard-deviation increase in Surp1 increased the odds of overt that by approximately 39%, OR = 1.39.

Peak surprisal also significantly increased complementizer realization in the reference condition, namely spoken production, $b = 0.272$, SE = 0.072, $z = 3.78$, $p < .001$. A one-standard-deviation increase in Peak surprisal corresponded to an odds ratio of 1.31.

In contrast, Mean surprisal showed a significant negative coefficient, $b = -0.467$, SE = 0.045, $z = -10.33$, $p < .001$. This coefficient represents the partial effect of Mean surprisal after clause-initial and Peak surprisal were held constant. It therefore indicates that, among clauses with comparable initial and peak values, a higher average surprisal level was associated with a lower probability of overt that.

Production mode also had a significant effect. Written responses were more likely to contain overt that than spoken responses, $b = 1.022$, SE = 0.102, $z = 10.05$, $p < .001$, corresponding to an odds ratio of 2.78.

Most importantly, the interaction between Peak surprisal and written mode was significant, $b = 0.265$, SE = 0.085, $z = 3.14$, $p = .002$. Thus, the positive effect of Peak surprisal was stronger in writing than in speech.

Predictor	Estimate	SE	z	p	Odds ratio
Intercept	-2.433	0.147	-16.55	< .001	0.09
Surp1	0.331	0.038	8.63	< .001	1.39
Mean surprisal	-0.467	0.045	-10.33	< .001	0.63
Peak surprisal	0.272	0.072	3.78	< .001	1.31
Written mode	1.022	0.102	10.05	< .001	2.78
L1: Japanese	0.896	0.098	9.19	< .001	2.45
L1: Korean	0.525	0.119	4.43	< .001	1.69
Proficiency: B1_1	0.141	0.108	1.31	.191	1.15
Proficiency: B1_2	0.077	0.124	0.63	.531	1.08
Proficiency: B2	-0.252	0.16	-1.58	.114	0.78
Peak surprisal × Written mode	0.265	0.085	3.14	.002	1.3

Note. Reference levels were Spoken, Chinese, and A2. All continuous surprisal predictors were standardized.

Table 2: Fixed effects of the best-fitting GLMM predicting overt complementizer realization

The Japanese and Korean groups showed higher complementizer-realization probabilities than the Chinese reference group. No proficiency-level contrast reached statistical significance.

4.3 Predicted Probability of Overt Complementizer Realization

Figure 1 illustrates the predicted probability of overt that across the observed range of standardized Peak surprisal in spoken and written production.

In both modes, the probability of overt complementizer realization increased as Peak surprisal became larger. However, the slope was considerably steeper in writing. In the spoken data, the predicted probability increased from approximately .05 at two standard deviations below the mean to approximately .13 at two standard deviations above the mean. In the written data, the corresponding probability increased from approximately .08 to more than .40.

This pattern visualizes the significant interaction between Peak surprisal and production mode. Peak surprisal therefore predicted overt that in both modes, but its effect was substantially stronger in written production.

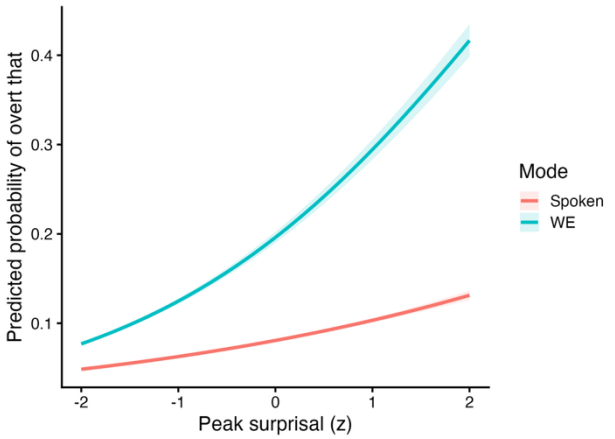

Figure 1: Predicted probability of overt that as a function of Peak surprisal and production mode.

4.4 Comparison Across Information Profiles

The model-comparison results showed that Peak surprisal was the best-performing individual predictor. Its AIC was lower than those of Surp1, Total surprisal, and Mean surprisal.

Nevertheless, the combined models performed much better than all single-predictor models. This indicates that complementizer realization cannot be reduced to one information-density profile alone. Clause-initial surprisal, average information density, and localized peak surprisal contributed independently to the prediction of overt *that*.

For this reason, the model-comparison figure should display ΔAIC rather than raw AIC values. Raw AIC values fall within a relatively narrow numerical range, causing the differences to appear visually compressed (Table 3).

Model	ΔAIC
Combined × Mode	0
Combined	7.63
Peak	162.61
Initial	168.12
Total	215.85
Mean	235.36

Table 3: Differences in AIC across models predicting overt complementizer realization.

4.5 Spoken Data

4.5.1 Filled pauses

To investigate whether localized information peaks were reflected in speech planning, the spoken data were examined for filled pauses. The spoken subset contained 1,907 observations, and uh or um occurred in 24.4% of them.

Peak surprisal significantly predicted filled-pause occurrence, $b = 0.435$, SE = 0.088, $z = 4.93$, $p < .001$. A one-standard-deviation increase in Peak surprisal increased the odds of a filled pause by approximately 55%, OR = 1.55 (Table 4).

Predictor	Estimate	SE	z	p
Intercept	-1.403	0.619	-2.27	.023
Peak surprisal,	**0.435**	**0.088**	**4.93**	**< .001**
L1: Japanese	-7.19	0.988	-7.28	< .001
L1: Korean	-0.317	0.264	-1.2	.231
Proficiency: B1_1	0.611	0.631	0.97	.333
Proficiency: B1_2	0.418	0.578	0.72	.470
Proficiency: B2	-0.864	0.675	-1.28	.201

Table 4: GLMM predicting filled-pause occurrence

Figure 2 shows the observed filled-pause rate by Peak-surprisal quartile. The rate generally increased across quartiles, from approximately .15 in the lowest quartile to approximately .33 in the highest quartile.

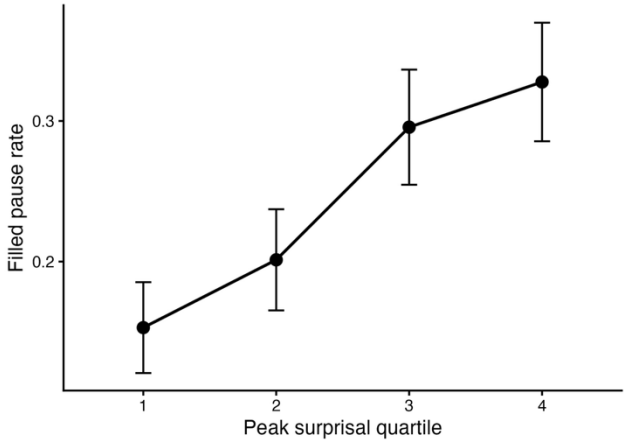

Figure 2: Observed filled-pause rate by Peak-surprisal quartile.

4.5.2 Silent pause duration

Pause duration before complement-clause onset was available for 476 spoken observations. The median pause duration was 200 ms, and the mean was 583.6 ms.

Peak surprisal did not significantly predict pause duration, b = 32.26 ms, SE = 43.81, t = 0.74, p = .462. The result remained nonsignificant when positive pause durations were log-transformed, b = 0.055, SE = 0.065, t = 0.86, p = .392.

4.5.3 Duration of overt *that*

The duration analysis included 497 overt instances of that. Their mean duration was 253.9 ms, and the median was 240 ms.

Peak surprisal did not significantly predict the duration of *that*, b = -4.27 ms, SE = 3.92, t = -1.09, p = .277.

Outcome	N	Estimate	SE	Test statistics	p
Pause before clause onset	476	32.26	43.81	t = 0.74	.462
Log-transformed positive pause	458	0.055	0.065	t = 0.86	.392
Duration of overt *that*	497	-4.27	3.92	t = -1.09	.277

Table 5: Effects of Peak surprisal on temporal measures

5. Discussion

5.1 Local Information Density and Complementizer Realization

The present study investigated which representation of information density best explains the optional realization of the English complementizer that. Previous work has repeatedly shown that optional complementizer realization is influenced by processing-related factors rather than by purely syntactic constraints (Ferreira & Dell, 2000; Jaeger, 2010). However, earlier studies have generally treated information density as a relatively global property of the upcoming clause, leaving open the question of which aspect of information density speakers are actually sensitive to. By directly comparing multiple surprisal-based representations within the same corpus, the present study provides evidence that different information-density profiles contribute differently to optional grammatical choice.

One important finding is that Peak surprisal was the strongest individual predictor of overt complementizer realization. This result extends previous UID research (Levy & Jaeger, 2007; Jaeger, 2010), which argued that speakers tend to distribute information so as to avoid excessive processing difficulty. Rather than contradicting UID, the present findings suggest a refinement of the theory. Speakers appear to be especially sensitive to localized informational bottlenecks within a clause rather than to its average information density. In other words, the probability of producing overt that increased when a highly unexpected word was anticipated later in the complement clause. This finding indicates that local peaks of processing difficulty play a particularly important role in grammatical optionality.

5.2 Multiple Information Profiles in Sentence Planning

At the same time, Peak surprisal alone did not provide the complete explanation. Although it yielded the best performance among the individual predictors, the combined model incorporating clause-initial surprisal, mean surprisal, and Peak surprisal substantially outperformed every single-measure model. This finding suggests that optional complementizer realization reflects multiple stages of linguistic planning. Clause-initial surprisal captures the processing demands immediately following complement-clause onset, mean surprisal reflects the overall informational profile of the clause, and Peak surprisal identifies localized processing bottlenecks. Together, these results suggest that speakers integrate both global and local information when making optional grammatical choices.

5.3 Information Density Across Production Modes

The interaction between Peak surprisal and production mode provides further insight into the planning process. The effect of Peak surprisal was significantly stronger in written production than in spoken production, suggesting that writers are better able to anticipate upcoming processing difficulty during sentence planning. In contrast, spontaneous speech is produced under stronger incremental constraints. Nevertheless, Peak surprisal significantly predicted the occurrence of filled pauses, indicating that localized processing difficulty also influences online speech production. Interestingly, Peak surprisal did not significantly predict silent pause duration or the acoustic duration of that. These findings imply that speakers respond to informational bottlenecks primarily by choosing whether to realize the complementizer, rather than by simply lengthening pauses or articulating the complementizer more slowly.

5.4 Methodological and Theoretical Implications

Methodologically, the present study also demonstrates the usefulness of modern neural language models for corpus-based investigations of grammatical optionality. As outlined above, GPT-2 was employed not as a theory of language production but as a practical tool for estimating the surprisal values through which UID's theoretical predictions were tested. This approach makes it possible to operationalize information density quantitatively and to compare alternative theoretical representations using authentic production data, extending corpus-linguistic methodology beyond earlier hand-coded predictability measures.

In summary, these findings motivate what we call the Peak Information Density Hypothesis (PIDH): whereas traditional UID accounts emphasize the regulation of overall information density, PIDH proposes that optional grammatical markers are preferentially realized before localized peaks of surprisal that constitute temporary processing bottlenecks. It is important to note that the present study examined only English complement clauses. Whether the same information profile characterizes other optional grammatical phenomena, such as optional relative pronouns, remains an open question. Future research should therefore investigate whether Peak surprisal represents a general principle of grammatical optionality or whether different constructions exhibit construction-specific information-density profiles.

6. Conclusion

This study examined which representation of information density best accounts for the optional realization of the English complementizer that in learner English. Building on the Uniform Information Density framework, we compared multiple surprisal-based measures derived from GPT-2, including clause-initial, mean, total, and peak surprisal.

Three main findings emerged. First, among the individual information-density measures, Peak surprisal was the strongest predictor of overt complementizer realization. Second, the best overall model combined clause-initial surprisal, mean surprisal, and Peak surprisal, indicating that speakers are sensitive to multiple aspects of information density rather than to a single summary measure. Third, the effect of Peak surprisal was significantly stronger in written production than in spoken production, while analyses of spoken data showed that Peak surprisal predicted filled pauses but not silent pause duration or the acoustic duration of that. Together, these findings suggest that localized informational bottlenecks influence grammatical choice and online speech planning in different ways.

The present study contributes to research on grammatical optionality in three respects. Theoretically, it refines the Uniform Information Density account by demonstrating the importance of localized information peaks in complementizer realization. Methodologically, it illustrates how neural language models such as GPT-2 can be used to operationalize psycholinguistic constructs in large learner corpora. Empirically, it provides quantitative evidence that multiple representations of information density jointly shape naturally occurring optional grammatical choices.

The findings motivate the Peak Information Density Hypothesis (PIDH), which proposes that optional grammatical markers are preferentially realized before localized peaks of processing difficulty. Because the present study focused exclusively on English complement clauses, future research should examine whether similar information-density profiles characterize other forms of grammatical optionality, including optional relative pronouns and additional syntactic constructions. Such investigations will help determine whether Peak surprisal represents a general principle of grammatical planning or whether different constructions exhibit distinct information-management strategies.

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant Number 24K00081.

References

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

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001), pages 1–8, Pittsburgh, PA, USA, June. Association for Computational Linguistics.

Hawkins, J. A. (2004). Efficiency and Complexity in Grammars. Oxford University Press, Oxford.

Ishikawa, S. (2023). The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English. Routledge, London.

Jaeger, T. F. (2005). Optional that indicates production difficulty: Evidence from disfluencies. In Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop, pages 103–109, Aix-en-Provence, France, September.

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

Levelt, W. J. M. (1989). Speaking: From Intention to Articulation. MIT Press, Cambridge, MA.

Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106(3):1126–1177.

Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT Press, pp. 849--856.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9.