12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)
4-5 September 2025, Lisbon, Portugal

Euh... where do interpreters hesitate? An information-theoretic perspective on sentence-initial filler particles in simultaneous interpreting

Christina Pollkläsener[1], *Maria Kunilovskaya*[2]

[1]Language and Information Sciences, University of Hildesheim, Germany
[2]Language Science and Technology, Saarland University, Germany
pollklaesener@uni-hildesheim.de, maria.kunilovskaya@uni-saarland.de

Abstract

This study investigates the occurrence of sentence-initial filler particles (e.g. euh, hum) in simultaneously interpreted and original speeches using a bidirectional English-German corpus of European Parliament debates. We assume that sentence-initial filler particles indicate planning difficulties at the conceptual level, whereas sentence-medial filler particles mark hesitations over syntactic structure or lexical access. Since interpreters convey the source speech and do not plan their own message, we expect differences between interpreting and original speeches. We operationalise conceptual complexity as average word surprisal per sentence and local lexical or syntactic production problems as surprisal of the word following the filler particle. Our findings indicate that sentence-initial filler particles appear in sentences with higher conceptual complexity but are not well associated with local retrieval difficulty.
Index Terms: hesitation markers, filler particles, interpreting, cognitive load, conceptual planning

1. Introduction

Filler particles (FPs, e.g. euh or hum, also called filled pauses or hesitation markers) are a frequent phenomenon in spoken language. These particles are syntactically non-essential, void of propositional meaning and often reflect difficulties in linguistic planning and production [1].

These properties render them particularly interesting in the context of simultaneous interpreting, a complex task requiring concurrent comprehension of source language input and production of target language output. Previous research has shown that FPs are generally more frequent in interpreting compared to non-mediated (original) speech [2, 3]. Current research uses the overall number of FPs as an empirical indicator of cognitive load, linking it to various properties of the source and target texts that might contribute to the difficulty of the task (e.g. numbers, high delivery rate, high lexical density [2, 4]).

The position of FPs in the sentence and their immediate context were studied before in non-mediated (original) speech production [5, 6, 7, 8, 9]. In simultaneous interpreting, Wang and Li [10] is the only study to our knowledge that investigated whether FPs appear in initial or medial position of phrases. However, they did not explore underlying causes or predictors of FP position.

Even though FPs can theoretically occur anywhere in a sentence, previous research has identified some regularities. FPs typically appear at phrase, clause and utterance boundaries [5, 11]. Furthermore, early work already highlighted the probabilistic specificity of FPs' immediate contexts [11, 12]. This was corroborated by more recent studies that found that FPs are followed by words with high surprisal [7, 8].

Surprisal, a measure rooted in information theory [13], quantifies the (un)expectedness of a word based on its contextual probability. It correlates with cognitive effort, as evidenced by measures like reading times, eye fixations, and event-related brain potentials [14, 15, 16, 17]: low-surprisal words require less comprehension effort, while high-surprisal words are more demanding. Moreover, some language production theories propose that surprisal influences lexical pre-activation, affecting the ease of word retrieval [18, 19]. The measure of average sentence surprisal (AvSrp) grounded in computational models (as employed in the current study) has been used before as a measure of sentence information in translation studies [20] and in clinical linguistics [21], where it was shown to reflect both lexical and structural components of sentence information. AvSrp (derived from cloze-test data) was reported to have a significant positive relationship with Flesch–Kincaid-based text difficulty [22]. This evidence motivates the use of AvSrp as a reasonable measure of conceptual planning difficulties.

The influential speech production theory by Levelt [23] divides the speech production process into three stages. The first stage is conceptualisation, where a prelinguistic message is formed as a semantic representation of an event. Message-level encoding is followed by grammatical encoding, i.e. syntactic planning and lexical access (formulation). The last stage involves phonetic encoding and articulation.

Kosmala and Crible [9] found quantitative and qualitative differences between FPs in sentence-initial and sentence-medial positions in originally-authored speeches. They concluded that FPs in initial position have a more fluent quality, signalling discourse boundaries, whereas in medial position they are more disfluent, indicating retrieval difficulties. Swerts [5] established that FPs marking major discourse boundaries were segmentally and prosodically different compared to FPs in other positions.

These findings suggest a functional distinction between FPs in initial position and those in medial position: the former reflect conceptual planning, the latter reflect problems at the lexical or syntactic level. This study tests whether this holds in interpreting.

Higher average surprisal values indicate a sentence's overall unpredictability. This can be treated as an indicator of sentences that are difficult to plan on the conceptual level. In contrast, high next-word surprisal signals local production difficulty. Building on this distinction, the present study aims to explore the link between initial FPs and conceptual planning difficulties, using AvSrp as the main measure of conceptual complexity and next-word surprisal as an indicator of local structuring or retrieval difficulties. We also track differences between the two text types (simultaneous interpreting and original speech) to test the hypothesis that conceptual planning in interpreting is less of an issue compared to original speech.

10.21437/DiSS.2025-19

Table 1: *Data parameters (de = German, en = English, org = originals, si = simultaneous interpreting)*

lang	ttype	docs	sents	tokens	total FPs	initial FPs
de	org	165	3,217	64K	604	87
en	si	165	3,673	63K	2,340	216
en	org	137	3,692	71K	1,196	248
de	si	137	3,293	64K	3,324	195

2. Methodology

2.1. Data

The data used in this study are drawn from the German and English subcorpora of EPIC-UdS [24]. They contain manual transcriptions of speeches held at the European Parliament between 2008 and 2013 by English and German native-speaking Members of Parliament, along with their respective simultaneous interpretations into German and English. The transcribers were instructed to include disfluencies such as filler particles, truncations and repetitions. The relevant metadata specifies speakers' and interpreters' identities.[1]

Table 1 summarises the quantitative parameters of the data, including total counts of FPs and those occurring in sentence-initial position, which are the focus of this study.

2.2. Surprisal values

Surprisal can be estimated either from non-contextualised probabilities based on relative frequencies in a corpus or from contextualised probabilities produced by a computational language model trained on large text data (e.g., n-gram models, LSTMs, or Transformers such as LLaMA and GPT). It is calculated as the negative base-2 logarithm of a token's probability, given the preceding context:

$$S(w_i) = -log2(P(w_i|w_1, w_2, ..., w_{i-1})) \quad (1)$$

where P is the probability of the word w_i and $w_1, w_2, ..., w_{i-1}$ is its preceding context.

In our data, surprisal was estimated per word, using language-specific pre-trained GPT-2 models with the vocabulary size of 50 K tokens (English [25], German [26]). We used GPT-2-small, as it fits reading times substantially better than larger GPT-2 variants (e.g., GPT-2-XL) or newer models (e.g., GPT-3) [27]. Prior to surprisal indexing, the data was parsed with Stanza [28]. The surprisal values were calculated for word-tokens as defined by the parser by aggregating log probabilities of the constituent subwords. Hyphenated words were treated as one word-token.

FPs were removed before parsing and surprisal annotation, and were reinserted into the corpus afterwards at their respective indices. Including them would have affected AvSrp and surprisal of words following the FPs (i.e., next-token surprisal), since they would have been treated as part of the preceding context for those words. Extracting values for sentence-initial tokens is challenging because these tokens lack preceding context for probability estimation. To address this, a beginning-of-sentence token (<bos>) was added to the model's vocabulary and prepended to each sentence. This allows the model to estimate the probability of a token to open a sentence.

[1] A student assistant was asked to identify individual interpreters by listening to interpreters' voices.

2.3. Modelling and Statistical Analysis

We used generalised linear mixed models (GLMMs) to analyse the occurrence of initial FPs. A document-level analysis of FP distribution between sentence-initial and sentence-medial positions was conducted using Poisson regression. For binary outcomes in sentence- and word-level experiments (presence or absence of initial FPs and medial FP or medial regular word, respectively), we applied logistic regression.

The regression models and statistical analyses were implemented in the R environment [29] using car [30], ggplot2 [31], emmeans [32] and lme4 [33] packages.

We fitted separate regression models for the German and the English data. The rationale was that while surprisal for both languages was derived from comparable language models, they were not identical. The probabilities from these models come from two distinct distributions due to differences in training corpora and in morphological structures between the two languages. A single regression model for both languages would have treated these values as coming from a single source.

Speaker or interpreter identity ("speaker_name") and document identity ("doc_id") were included as random intercepts in all models. These random variables take into account the variation in the frequency of FPs across speakers and documents. In this setup, the effect of the main predictors (e.g., AvSrp) is assumed to be the same across speakers/documents.

The key predictors used in the sentence-level and/or word-level experiments are outlined below:

ttypeC: text type (interpreting or original), sum-coded, i.e. the reference level (0) is "centred" halfway between the two categories (interpreting coded as +1, originals coded as -1),

next_srp: the surprisal of the immediately following word; for initial FPs, it is the first word in a sentence (recall that initial FPs were not taken into account when estimating surprisal),

AvSrp average sentence surprisal, i.e. mean surprisal across the sentence tokens, excluding punctuation,

AvDD average dependency distance, a well-established alternative measure of sentence comprehension complexity with implications for cognitive costs of processing [34, 35, 36],

medialFP the count of the sentence-medial FPs,

sentence_length sentence length in Stanza tokens, excluding FPs and punctuation.

The document-level experiment used two main categorical predictors, which are dummy-coded: text type (originals as reference category, coded as 0, interpreting coded as 1) and position of FP (initial as reference category, coded as 0, medial coded as 1) to predict the number of FP in each position per document. Additionally, we coded interactions between text type and key predictors: next_srp, AvSrp and AvDD. All continuous variables were z-transformed to make them comparable to each other, and to facilitate interpretability and convergence [37].

3. Results

3.1. Initial and medial FPs in originals vs. interpreting

In this document-level analysis, we examine the distribution of initial and medial FPs in original and interpreted speeches, assuming that there are differences between the text types and positions. We fit two Poisson mixed-effects models with counts of FPs per speech as the dependent variable. The models included fixed effects for text type (original vs. interpreting), position (initial vs. medial), and their interaction:

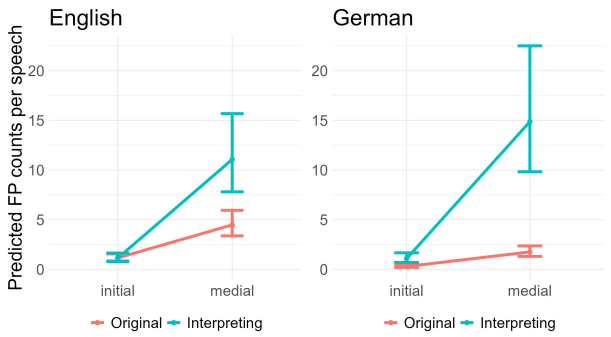

Figure 1: *Predicted FP counts per speech in the English (left) and German (right), grouped by text type and position.*

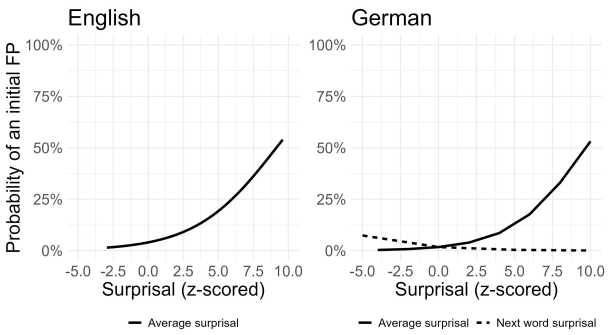

Figure 2: *Effect of sentence-average and next-word surprisal on probability of an initial FP per sentence in the English and German data.*

FP_count $\sim$ *ttype + position + ttype:position + offset(log(n_sentences)) + (1 | speaker_name) + (1 | doc_id).*

We fit the number of sentences per speech as an exposure variable, so that the model treats the number of sentences per speech as fixed (24 sentences per English speech, 21 sentences per German speech). We assessed overdispersion in the model using the check_overdispersion() function from the performance package [38]. Neither the English nor the German model show evidence for overdispersion (dispersion ratio = 0.886, $p > .979$ and ratio = 0.640, $p > 1$, respectively).

The English model reveals a significant interaction between position and text type ($\beta = 0.92$, SE = 0.10, $p < .001$), indicating that the difference between initial and medial FP counts is larger for interpreting compared to originals, as seen in Figure 1. Pairwise comparison of estimated marginal means further reveals that FP counts are significantly higher in medial compared to initial positions for both text types ($p < .001$). Comparing across text types, we see that in initial positions, there is no significant difference in predicted FP counts between originals and interpreting ($p > .999$), whereas in medial positions, FP counts are significantly higher in interpreting compared to originals ($p < .001$).

For the German data, we observe similar trends. The right-hand side of Figure 1 shows that there is a significant interaction between position and text type ($\beta = 0.81$, SE = 0.14, $p < .001$), indicating that the difference between initial and medial FP counts is more pronounced in interpreted compared to original documents. When examining the effect of position within each text type separately, the difference in counts between initial and medial positions is significant for both original and interpreted texts ($p < .001$). Additionally, comparisons across text types reveal significant differences between interpreting and originals for both FP counts in initial and medial positions ($p < .001$). This is unlike the English data, where initial position counts do not differ significantly between text types.

3.2. Initial FPs: Conceptual vs. local difficulties

After examining the overall distribution of initial and medial FPs in original and interpreted speeches, the analysis now turns to potential explanatory variables for FPs in sentence-initial position.

In this experiment, the sentence-level binary response variable of the logistic regression model indicates whether a sentence opens with a FP (1) or not (0). The predictors and coefficients of the model are listed in Table 2. Note that here,

Table 2: *Estimates and significance levels for predictors in English and German sentence-level models. Significance codes:* *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Predictor	English	German
(Intercept)	-3.168 ***	-4.036 ***
AvDD_z	-0.007	0.037
AvSrp_z	0.347 ***	0.415 ***
medialFP_z	0.109 *	0.089
next_srp_z	0.085	-0.300 ***
sentence_length_z	0.081	0.202 **
ttypeC	-0.102	0.514 **
ttypeC:AvDD_z	0.013	0.060
ttypeC:AvSrp_z	-0.033	0.062
ttypeC:next_srp_z	-0.096	0.001

next_srp_z means the surprisal of the word following the initial FP or the surprisal of the initial word of the sentence without an opening FP. To assess potential collinearity issues, we examine the models' variance inflation factors (VIFs). In both English and German datasets, all VIF values are below 2, indicating that collinearity is not a concern. We focus our discussion of results on the main predictors.

As can be seen in Table 2, for the English model, AvSrp emerges as the strongest significant predictor for FPs in sentence-initial position. The higher AvSrp of a sentence, the higher the probability of a FP at the sentence-initial position (cf. left side Figure 2). In contrast, next-word surprisal has no significant effect. Text type is not a significant predictor, confirming our observations in Section 3.1: there is no difference between interpreting and originals for initial FPs. There are no significant interactions, suggesting that AvSrp's effect on initial FP occurrence does not differ across text types.

The German model shows some differences compared to the English model. As reflected in the results in Section 3.1, the significant, positive estimate for text type indicates that there are more initial FPs in interpreting than in originals. The significant effect of the next-word surprisal is depicted in Figure 2 (cf. right side). The relation has an unexpected trend: the higher the surprisal of the first word in the sentence, the lower the probability of the sentence having an initial FP. AvSrp has the strongest significant effect among the continuous predictors. Like in the English data, the higher AvSrp, the more likely the sentence begins with an initial FP.

Table 3: *Estimates and significance levels for predictors in English and German word-level models. Significance codes:* *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Predictor	English	German
(Intercept)	-4.024 ***	-4.146 ***
AvSrp_z	0.104 ***	0.112 ***
next_srp_z	0.383 ***	0.306 ***
ttypeC	0.567 ***	1.073 ***
ttypeC:AvSrp_z	-0.047 *	-0.021
ttypeC:next_srp_z	0.093 ***	-0.024

Similar to what we see in the English model, there are no significant interactions in the German data, indicating that there is no difference between originals and interpreting in the effects of AvSrp and next-word surprisal on the occurrence of a FP at the beginning of a sentence.

3.3. Medial FPs: Conceptual vs. local difficulties

This experiment aims to test whether sentence-medial FPs are more associated with conceptual planning or with local structuring problems. It partly reproduces our previous study [39], using the predictors accessible in the scope of the current study. Here, we fit a logistic regression model to a word-level binary response variable distinguishing between sentence-medial FPs and regular words in the corpus. Predictors and results of the models appear in Table 3. In both the English and German datasets, VIF values for AvSrp and next-word surprisal are below 2, indicating that collinearity is not a concern.

The results for text type confirm the findings in Section 3.1, showing that medial FPs are more likely in interpreting in both languages, independent of surprisal. Higher surprisal (both next-word and sentence-average) is associated with a higher likelihood of a word to be a sentence-medial FP. However, in both languages, FPs inside the sentence are more associated with next-word surprisal than with AvSrp. These surprisal effects seem to be the same in interpreting as in originals in German (no significant interaction between either of the surprisal measures and the text type). Importantly, in English, next-word surprisal is higher in interpreting than in originals (as shown by the interaction between text type and next-word surprisal). This could indicate that interpreters face more local difficulties and source language interference when rendering the German source speech into English, compared to the local encoding difficulties that the original English speakers face. There is also an interaction between AvSrp and text type in English: sentence-average surprisal for medial FPs has a significantly lower effect in interpreting than in original speeches. If AvSrp represents conceptual difficulty, this result might mean that interpreters into English struggle less with conceptual planning than original speakers do.

The systemic analysis of the variability of the speakers and documents as random intercepts indicates that the variance magnitudes are stable across the three experiments and languages (not given here due to space constraints). Speakers contribute more random variability than documents (i.e., context). Speakers' individual preferences had a considerable impact on the predictions, with SD averaging at 0.91 for English and 1.06 for German. German shows greater variability in both grouping factors than English, pointing to less homogeneity in how speakers and documents influence the occurrence of FPs.

4. Discussion and conclusion

The present study highlights a functional distinction between initial and medial FPs. In the sentence-level setup, AvSrp was the only significant predictor of the presence of a FPs in the beginning of the sentence in English. In German, the competing predictor, next-word surprisal, had a significant effect but in the opposite direction, showing that initial FPs were less likely if the first word in the sentence had higher surprisal. It strongly supports the assumption that initial FP are not about local difficulties. The word-level experiment tested the same predictors on the occurrence of medial FPs. Here, although both AvSrp and next-word surprisal significantly predicted the outcome, the effect of next-word surprisal was approximately three times stronger than that of AvSrp. Additionally, the total number of FPs in a sentence, another indicator of sentence complexity, emerged as a relatively good predictor of the occurrence of an initial FP. These findings confirm that initial FPs capture hesitations about conceptual planning to a greater extent than FPs in the sentence-medial positions, which are more associated with local processing difficulties.

Our results also establish AvSrp as a possible measure of conceptual planning difficulties. It proved to be a reasonably strong predictor of initial FPs in comparison with AvDD as an alternative measure of sentence complexity that did not return significant results in any of our languages. Regarding overall differences across text types, we found differences in the frequency of FPs per position, but not in the tendencies for explanatory predictors. As expected, interpreting had significantly more FPs than original speech, including in initial position (for German only). Contrary to our initial expectations, initial and medial filler particles serve similar functions in both original and interpreted speech. In interpreted English, medial filler particles appear to be even more strongly associated with local production difficulties than with conceptual planning, as evidenced by the interaction effects between text type and each surprisal measure observed in Study 3.

There are several noteworthy limitations to this study. First, while AvSrp shows promise as an indicator of conceptual complexity, its validity should be tested in controlled experimental settings. Second, our operationalisation of 'sentence' may not align fully with definitions of 'utterance' or 'inter-pausal units' used in other studies, limiting comparability. On a more abstract level, conventional sentence structures might not align well with the actual scope of message planning. Finally, this study did not explore other predictors additional and/or alternative to surprisal (except dependency distance). We cannot rule out the possibility that those unknowns can also explain the role of filler particles in various positions as signals of conceptual or local planning difficulties.

5. Acknowledgments

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1102 Information Density and Linguistic Encoding, Project-ID 232722074.

6. References

[1] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[2] K. Plevoets and B. Defrancq, "The cognitive load of interpreters in the European Parliament: A corpus-based study of predictors

for the disfluency uh (m)," *Interpreting*, vol. 20, no. 1, pp. 1–28, 2018.

[3] A. Chmiel, D. Koržinek, M. Kajzer-Wietrzny, P. Janikowski, D. Jakubowski, and D. Polakowska, "Fluency parameters in the Polish Interpreting Corpus (PINC)," *Mediated discourse at the European Parliament: Empirical investigations*, pp. 63–91, 2022.

[4] M. Kajzer-Wietrzny, I. Ivaska, and A. Ferraresi, "Fluency in rendering numbers in simultaneous interpreting," *Interpreting*, vol. 26, no. 1, pp. 1–23, 2024.

[5] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of pragmatics*, vol. 30, no. 4, pp. 485–496, 1998.

[6] R. Rose, "Silent and filled pauses and speech planning in first and second language production," *Proceedings of DiSS*, vol. 2017, pp. 49–52, 2017.

[7] J. Zámečník, "Disfluency prediction in natural spoken language," Ph.D. dissertation, Dissertation, Universität Freiburg, 2019, 2019.

[8] S. Dammalapati, R. Rajkumar, S. Ranjan, and S. Agarwal, "Effects of duration, locality, and surprisal in speech disfluency prediction in English spontaneous speech," in *Proceedings of the Society for Computation in Linguistics 2021*, 2021, pp. 91–101.

[9] L. Kosmala and L. Crible, "The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence," *Language and Speech*, vol. 65, no. 1, pp. 216–239, 2022.

[10] B. Wang and T. Li, "An empirical study of pauses in Chinese-English simultaneous interpreting," *Perspectives*, vol. 23, no. 1, pp. 124–142, 2015.

[11] F. Goldman-Eisler, *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press, 1968.

[12] G. W. Beattie and B. L. Butterworth, "Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech," *Language and speech*, vol. 22, no. 3, pp. 201–211, 1979.

[13] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[14] V. Demberg and F. Keller, "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity," *Cognition*, vol. 109, no. 2, pp. 193–210, 2008.

[15] M. Kutas, K. A. DeLong, and N. J. Smith, "A look around at what lies ahead: Prediction and predictability in language processing," *Predictions in the brain: Using our past to generate a future*, vol. 190207, no. 10.1093, 2011.

[16] N. J. Smith and R. Levy, "The effect of word predictability on reading time is logarithmic," *Cognition*, vol. 128, no. 3, pp. 302–319, 2013.

[17] J. A. Michaelov, M. D. Bardolph, C. K. Van Petten, B. K. Bergen, and S. Coulson, "Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects," *Neurobiology of Language*, vol. 5, no. 1, pp. 107–135, Apr. 2024.

[18] G. R. Kuperberg and T. F. Jaeger, "What do we mean by prediction in language comprehension?" *Language, cognition and neuroscience*, vol. 31, no. 1, pp. 32–59, 2016.

[19] F. Huettig, J. Audring, and R. Jackendoff, "A parallel architecture perspective on pre-activation and prediction in language processing," *Cognition*, vol. 224, p. 105050, 2022.

[20] M. Kunilovskaya, H. Przybyl, E. Teich, and E. Lapshinova-Koltunski, "Simultaneous interpreting as a noisy channel: How much information gets through," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria: INCOMA Ltd., 2023-09-04/2023-09-06, pp. 608–618.

[21] N. Rezaii, J. Michaelov, S. Josephy-Hernandez, B. Ren, D. Hochberg, M. Quimby, and B. C. Dickerson, "Measuring Sentence Information via Surprisal: Theoretical and Clinical Implications in Nonfluent Aphasia," *Annals of Neurology*, vol. 94, no. 4, pp. 647–657, 2023.

[22] M. W. Lowder, W. Choi, F. Ferreira, and J. M. Henderson, "Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction," *Cognitive Science*, vol. 42, no. S4, pp. 1166–1183, 2018.

[23] W. J. M. Levelt, *Speaking: From Intention to Articulation*. Cambridge, Mass: The MIT Press, 1989. [Online]. Available: https://doi.org/10.7551/mitpress/6393.001.0001

[24] H. Przybyl, E. Lapshinova-Koltunski, K. Menzel, S. Fischer, and E. Teich, "EPIC-UdS - creation and applications of a simultaneous interpreting corpus," in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. Marseille, France: ELDA, 20-25 June 2022, pp. 1193–1200.

[25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI*, 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[26] S. Schweter, "German GPT-2 model," *Zenodo*, November 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4275046

[27] B.-D. Oh and W. Schuler, "Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?" *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 336–350, 03 2023. [Online]. Available: https://doi.org/10.1162/tacl_a_00548

[28] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[29] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024. [Online]. Available: https://www.R-project.org/

[30] J. Fox and S. Weisberg, *An R companion to applied regression*. Sage publications, 2018.

[31] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[32] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2024, r package version 1.10.4. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[33] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[34] Y. Jing and H. Liu, "Mean Hierarchical Distance Augmenting Mean Dependency Distance," in *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, J. Nivre and E. Hajicova, Eds., 2015, pp. 161–170. [Online]. Available: http://ufal.mff.cuni.cz/pcedt2.0/

[35] L. Fan and Y. Jiang, "Can dependency distance and direction be used to differentiate translational language from native language?" *Lingua*, vol. 224, pp. 51–59, 2019. [Online]. Available: https://doi.org/10.1016/j.lingua.2019.03.004

[36] W. Xu and R. Futrell, "Syntactic dependency length shaped by strategic memory allocation," in *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1–9. [Online]. Available: https://aclanthology.org/2024.sigtyp-1.1/

[37] B. Winter, *Statistics for linguists: An introduction using R*. Routledge, 2019.

[38] D. Lüdecke, M. S. Ben-Shachar, I. Patil, P. Waggoner, and D. Makowski, "performance: An R package for assessment, comparison and testing of statistical models," *Journal of Open Source Software*, vol. 6, no. 60, p. 3139, 2021.

[39] C. Pollkläsener, M. Kunilovskaya, and E. Teich, "Surprisal explains the occurrence of filler particles in simultaneous interpreting," *SKASE Journal of Translation and Interpretation*, 2025, under review.