Determinants of Hesitations and Repetitions in Hindi Spontaneous Speech

Eashani Sharma* **and Ishita Arun*** **and Samar Husain**
Indian Institute of Technology Delhi
eashani.sharma@hss.iitd.ac.in
huz248352@iitd.ac.in
samar@hss.iitd.ac.in

Abstract

This study investigates the factors that predict disfluencies in Hindi spontaneous speech. In particular, we probe the influence of lexical, syntactic, phonological, and prosodic factors on two kinds of disfluencies, namely, hesitations and repetitions. These disfluencies are probed through both the nature of linguistic factors as well as through the source (preceding vs. following word) of these factors. Our results show that hesitations and repetitions pattern differently during spontaneous speech. Hesitations increase due to lexical, syntactic, as well as articulatory features from both preceding and following words. On the other hand, repetitions arise mainly due to lexical and articulatory factors of the upcoming word. Further, while previous research (e.g., Bell et al., 2009; Dammalapati et al., 2021) on English highlights the importance of upcoming difficulty on disfluencies, our results suggest that previously encountered difficulties can also lead to an increase in disfluencies. This suggests that language typology (SVO vs SOV) can play a critical role in determining the planning process and thereby affecting the distribution of disfluencies in a language. Together, these findings highlight the need for increased cross-linguistic research to understand the nature of incrementality and monitoring of the production system cross-linguistically.

1 Introduction

The conversion of thought to speech is complex and involves cognitive, linguistic, and motor faculties. It is commonly assumed that one transforms the conceptual message into a linguistic formation, assigning the lexical counterparts of the concepts and incorporating these lexical items into a syntactic structure. The output of this formulation then undergoes phonological encoding, followed by motor planning for speech output (Garrett, 1975; Bock and Levelt, 1994). While much of the speech is error-free, there is also a substantial amount of disfluencies, pauses, and repairs (Hockett, 1967; Nooteboom, 1980; Fromkin, 1971, 1973; Brennan and Williams, 1995; Lickley, 2017). Any interference or difficulty in the production process, such as delays in planning (Clark and Fox Tree, 2002), lexical retrieval, and phonological difficulties (Fraundorf and Watson, 2013; Levelt, 1983; Goldman-Eisler, 1961), a change in discourse status (Arnold et al., 2003), processing load (Bortfeld et al., 2001), or repairs (Levelt, 1983; Postma and Kolk, 1993; Blackmer and Mitton, 1991), can lead to disfluencies. In fact, it has been suggested that around 6-10% words in every 100 words are disfluent (Fox Tree, 1995).

Given that disfluencies tend to have no relation to the message's central proposition (Bortfeld et al., 2001; Fraundorf and Watson, 2013; Cossavella and Cevasco, 2021; Shriberg, 1994), they have traditionally been excluded from dedicated investigations. Under this view, they reflect production difficulties rather than linguistic meaning (Levelt, 1983; Goldman-Eisler, 1961). However, recent evidence suggests that fillers like *'Uh'* and *'Um'* hold lexical status (Clark and Fox Tree, 2002; Arnold et al., 2007). The multi-stage nature of speech production and the distributed occurrence of disfluencies in natural speech make it difficult to isolate their underlying causes: do they arise from lexical, syntactic, or any other processing bottlenecks?

Some attempts have been made in English to investigate this using corpus-based methods. The existing literature on the Switchboard corpus has revealed the following distributional patterns (particularly for filled pauses): disfluencies are likely to occur before less predictable words (Arnold et al., 2007; Sen, 2020). Words preceding disfluencies display low lexical surprisal and syntactic complexity, while words following disfluencies exhibit high lexical surprisal and syntactic complexity

*Equal contribution.

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 59–71

(Dammalapati et al., 2021). On the prosodic front, words neighboring disfluencies are longer or show elongated duration (Dammalapati et al., 2021), often accompanied by a prosodic flattening (Zayats and Ostendorf, 2019; Székely et al., 2017), indicating increased processing load (Ferreira, 1993). While these lexical, syntactic, and prosodic factors target different levels of sentence processing, they fall short in two respects: they do not account for typologically different languages where processing strategies may differ, and they largely overlook the role of phonological planning.

We extend previous work by systematically addressing the causes of disfluencies at different levels of sentence processing in Hindi, a Subject-Object-Verb (SOV) language. Disfluencies manifest in various forms: hesitations, repetitions, self-corrections, false starts, and reparanda (Shriberg, 1994; Fox Tree, 1995, 2001). We focus specifically on two types: hesitations and repetitions. Hesitations are interruptions in speech marked by either empty pauses (silent breaks) or filled pauses (such as *'Uh'* and *'Um'*) (Fox Tree, 2001; Clark and Fox Tree, 2002). Because it is difficult to reliably differentiate between fluent and disfluent empty pauses (Fox Tree, 1995), we focus our analysis on filled pauses and refer to them as *hesitations* hereafter. Repetitions involve the reiteration of a word or phrase (Fox Tree, 2001; Clark and Fox Tree, 2002; Levelt, 1983).

To identify the locus of difficulty during sentence production, we examine both preceding and following words across four processing levels using the following features:

1. **Lexical**: frequency, semantic similarity, and bigram probability
2. **Structural**: integration cost and surprisal
3. **Phonological**: word length[1] and phonological complexity
4. **Prosodic**: duration, pitch, and intensity

We address predictability at the lexical and syntactic levels using surprisal (Hale, 2001; Levy, 2008), semantic similarity (Miller and Charles, 1991; Roland et al., 2012), and bigram probabilities (Jurafsky and Martin, 2013); structural complexity with integration cost (Gibson et al., 2000); and phonological complexity with word length and phonotactic complexity (Vitevitch and Luce, 1999; Vitevitch et al., 2004; Celata, 2020). The associated cognitive effort is informed by prosodic analysis on a subset of the data. This approach offers a comprehensive understanding of disfluencies.

The paper is organized as follows: Section 2 introduces our dataset and discusses analyses at different linguistic levels, along with our predictions. Section 3 presents the results of the mixed-effects models for hesitations and repetitions, including prosodic predictors on the hesitation data subset. Section 4 discusses our findings and their implications. Section 5 concludes the paper.

2 Methods

2.1 Dataset

We use the IIT Delhi Hindi Dialog Corpus (Pareek et al., 2025), a collection of 60 unscripted spoken telephonic conversations from the Hindi segment of the CALLFRIEND Project (Canavan and Zipperlen, 1996). This gold standard corpus contains 39,244 sentences that have been transcribed and tagged for universal Parts of Speech (POS) categories and dependency relations.

Additionally, different types of disfluencies are encoded at the token level in the CoNLL-U format using specific tags stored in the miscellaneous (MISC) column. We focus on two such tags, specifically, filled pauses that are marked with a `Hesitation` tag, and repetitions that are marked with a `Disfluency` tag on the corresponding token (see Figure 1). In our analysis, we refer to these as hesitations and repetitions, respectively. We extracted all sentences containing at least one of these tags, yielding 2,289 utterances with hesitations (5.8%) and 1,379 utterances with repetitions (3.6%).

2.2 Analysis

The analysis proceeded in several stages. First, sentences containing hesitations and repetitions were extracted using the CONLL-U parser in Python (Buchholz and Marsi, 2006). For each target sentence, we identified the words immediately preceding and following hesitations or repetitions and extracted their linguistic features. We also extracted control sentences without speech errors, matched for sentence length, to serve as a baseline for fluent speech. Each target sentence was paired with one control sentence such that the control token occurred at the same index position as the hesitation

[1]Word length, measured as syllable count, was categorized as a phonological predictor, as a greater number of syllables imposes greater phonological planning demands.

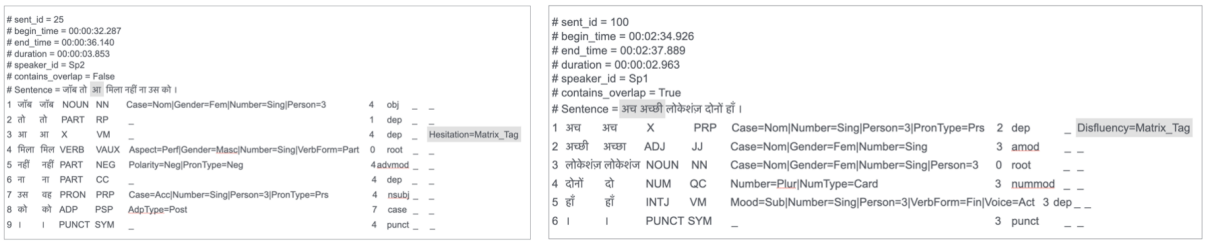

Figure 1: Example sentences from the IIT Delhi Hindi Dialogue Corpus in CoNLL-U format with transcription-level annotations and non-linguistic metadata. Hesitation (left): the hesitation token and its corresponding annotation in the miscellaneous column are highlighted in gray. Repetition (right): the disfluent token and its corresponding annotation are highlighted in gray.

or repetition token. Finally, we fitted mixed-effects logistic regression models using the lme4 package in R (Bates et al., 2015).

2.2.1 Linguistic features

Since disfluencies are widely understood to reflect difficulties at different stages of speech production, including lexical retrieval (Levelt, 1983, 1989), syntactic planning (Dammalapati et al., 2021; Bock and Levelt, 1994; Scontras et al., 2015), and articulation (Bell et al., 2009; Székely et al., 2017), difficulty at any stage can surface locally in the speech signal (Clark and Fox Tree, 2002; Levelt, 1983). Motivated by this, we investigated the following linguistic features in words immediately preceding or following hesitations and repetitions, as well as matched control positions, to capture production difficulty across multiple processing levels.

- **Word Frequency:** An estimate of how frequently a word occurs in the corpus. To normalize this value, the raw frequency count was Zipf-transformed.
- **Surprisal:** A measure of the unexpectedness of a word in a context, computed as the negative logarithm of the probability of the word given the preceding context (Hale, 2001; Levy, 2008).
- **Semantic similarity:** An estimate of how semantically related a word is to its preceding context, computed as the cosine similarity between the word's embedding and a vector representation of the preceding context (Miller and Charles, 1991; Roland et al., 2012).
- **Integration cost:** A measure of syntactic complexity, computed as the sum of linear distances between the target word and each of its linearly preceding head/dependents, with an additional baseline cost of 1 (Gibson et al., 2000).
- **Bigram probability:** The probability of a word given the preceding word. When a bigram was unavailable, a fallback to unigram probabilities was applied (Jurafsky and Martin, 2013).
- **Word length:** The number of syllables in a word.
- **Phonological complexity:** A measure of phonotactic complexity, computed as the probability of a consonant–vowel (CV) sequence given the preceding CV patterns found in the words of the corpus.

Details of the computational implementation of each feature are provided in Appendix A.

After feature extraction, we performed several pre-processing steps in R (R Core Team, 2025). We removed duplicate sentences and cases with missing values for any predictor. We also excluded sentences with hesitations or repetitions in sentence-initial or sentence-final positions, as these do not permit analysis of both preceding and following lexical contexts.

For hesitations, we obtained 2,095 unique sentences from the corpus. After excluding 673 sentences with sentence-initial hesitations (32%) and 97 with sentence-final hesitations (4.6%), and removing cases with missing values, we retained 897 sentences containing hesitations. These were paired with 897 matched control sentences, yielding a final dataset of 1,794 sentences.

For repetitions, we obtained 1,379 unique sentences. After excluding 329 sentences with repetitions at the sentence onset (23.86%), 23 at the sentence end (1.67%), and removing rows with missing values, we retained 780 sentences. These were paired with 780 matched control sentences, yielding a final dataset of 1,560 sentences.

We fitted separate generalized linear mixed-effects (*glmer*) models for hesitations and rep-

etitions, with a binary outcome variable (presence/absence of disfluency) as the dependent variable. Both models included linguistic features with only weak to moderate correlations[2] of the preceding and following words as fixed-effect predictors, and conversation ID[3] as a random intercept.

The hesitation model included an additional interaction term between preceding word frequency and length. The repetition model excluded word length[4] as a predictor and instead included an interaction between following word frequency and semantic similarity. Model selection followed standard recommendations for generalized linear mixed-effects models, including convergence checks and the comparison of nested random-effects structures whenever necessary (Winter, 2019; Barr et al., 2013).

$$\begin{aligned}\texttt{Hesitation} \sim\ & \texttt{WordFreq}_{\text{prev}} + \texttt{WordFreq}_{\text{next}} + \texttt{IC}_{\text{prev}} \\ & + \texttt{IC}_{\text{next}} + \texttt{SemSim}_{\text{prev}} + \texttt{SemSim}_{\text{next}} \\ & + \texttt{Surprisal}_{\text{prev}} + \texttt{Surprisal}_{\text{next}} \\ & + \texttt{BiProb}_{\text{prev}} + \texttt{BiProb}_{\text{next}} \\ & + \texttt{WordLen}_{\text{prev}} + \texttt{WordLen}_{\text{next}} \\ & + \texttt{PhonComp}_{\text{prev}} + \texttt{PhonComp}_{\text{next}} \\ & + \texttt{WordFreq}_{\text{prev}} * \texttt{WordLen}_{\text{prev}} \\ & + (1 \mid \texttt{ConversationID})\end{aligned}$$

$$\begin{aligned}\texttt{Repetition} \sim\ & \texttt{WordFreq}_{\text{prev}} + \texttt{WordFreq}_{\text{next}} + \texttt{IC}_{\text{prev}} \\ & + \texttt{IC}_{\text{next}} + \texttt{SemSim}_{\text{prev}} + \texttt{SemSim}_{\text{next}} \\ & + \texttt{Surprisal}_{\text{prev}} + \texttt{Surprisal}_{\text{next}} \\ & + \texttt{BiProb}_{\text{prev}} + \texttt{BiProb}_{\text{next}} \\ & + \texttt{PhonComp}_{\text{prev}} + \texttt{PhonComp}_{\text{next}} \\ & + \texttt{WordFreq}_{\text{next}} * \texttt{SemSim}_{\text{next}} \\ & + (1 \mid \texttt{ConversationID})\end{aligned}$$

2.2.2 Prosodic features

To examine the prosodic context surrounding hesitations, we sampled 130 sentences with 6 to 16 tokens containing hesitations and 130 control sentences without any speech errors, matched for sentence length. For sentences without hesitation, we aligned the token positions with those of the hesitations to ensure the parallel extraction of the preceding and following words for comparative analysis. These sentences were annotated in PRAAT version 6.4.26 (Boersma, 2012) at the sentence and word levels and exported to TextGrid files. A custom PRAAT script was used to extract pitch, duration, and intensity measurements for the target sentences, as well as for the preceding and following words.

Statistical analysis was conducted at the sentence and word levels in R. First, we examined how sentence-level prosodic properties could predict the presence of hesitations by fitting a *glmer* model with a binary outcome variable indicating the presence or absence of hesitations. Fixed-effect predictors included sentence duration, average pitch, and average intensity, all standardized prior to modeling. To capture joint effects of temporal and prosodic variation, the model additionally included interactions between duration and average pitch, and between average pitch and sentence length. Speaker [5] was included as a random intercept.

$$\begin{aligned}\texttt{Hesitation} \sim\ & \texttt{Duration} + \texttt{Intensity} + \texttt{Pitch} \\ & + \texttt{Duration} * \texttt{Pitch} \\ & + \texttt{Pitch} * \texttt{SentenceLength} \\ & + (1 \mid \texttt{Speaker})\end{aligned}$$

For the word-level analysis, we investigated how the prosodic features in the immediate context varied as a function of hesitations. Sum contrasts were applied to two categorical predictors, hesitation (present vs. absent) and token position (preceding vs. following), to estimate their main effects and interaction. Separate linear mixed-effects models were fitted for duration, pitch, and intensity, with hesitation and token position specified as fixed-effect predictors. Speaker, gender[6], and sentence length were included as random intercepts to account for participant, demographic, and utterance-level variability. Model selection followed standard recommendations for linear mixed-effects models, including convergence diagnostics and comparisons of nested random-effects structures (Winter, 2019; Barr et al., 2013).

2.3 Predictions

Under the assumption that previous findings would hold cross-linguistically, we hypothesized that hesitations and repetitions would be more likely

[2] All continuous predictors were z-scored prior to model fitting. Phonological complexity was log-transformed before standardization to reduce skew.

[3] Conversation ID uniquely identifies each telephone conversation in the corpus (corresponding to the recording filename). It serves as a discourse-level random effect in our mixed-effects models to account for within-conversation variability.

[4] Word length was excluded as a predictor due to model non-convergence.

[5] Speaker denotes an individual participant in a recorded telephone conversation and serves as a random effect to capture participant-level variability.

[6] Speaker and gender information is available as metadata in the IIT Delhi Hindi Dialog corpus (Pareek et al., 2025).

when words preceding them have low lexical surprisal and integration costs, while words following them have higher surprisal and integration costs (Dammalapati et al., 2021; Arnold et al., 2003). Since speech errors are known to arise from an increase in planning load (Clark and Fox Tree, 2002; Levelt, 1983; Fraundorf and Watson, 2013; Arnold et al., 2003), we also predicted that hesitations and repetitions would be more likely to occur when words following them have lower word frequency and bigram probability, along with higher semantic similarity[7] and phonological complexity.

For prosodic features, we hypothesized that words preceding hesitations would have a lower pitch, while words following hesitations would have a higher pitch to mark restarts (Zayats and Ostendorf, 2019; Székely et al., 2017). Similarly, for duration, we hypothesized that words preceding hesitations would be longer due to increased planning load and retrieval delay, and that words following hesitations would be shorter as fluency is regained (Dammalapati et al., 2021). The average intensity surrounding hesitations was anticipated to differ along the same lines, typically decreasing before hesitations and rising after them.

3 Results

3.1 Linguistic predictors of hesitations

Hesitations were predicted by frequency ($p < .05$), integration cost ($p < .05$), surprisal ($p < .01$), and phonological complexity ($p < .01$) of the preceding word, along with integration cost ($p < .01$), semantic similarity ($p < .01$), word length ($p < .001$), and phonological complexity ($p < .01$) of the following word. Hesitations increased with higher surprisal, higher integration cost, and greater phonological complexity of the preceding word. In addition, hesitations increased with higher integration cost, increased semantic similarity, increased word length, and phonological complexity of the following word. A higher frequency of the preceding word reduced the likelihood of hesitations. A significant interaction between the frequency and length of the preceding word ($p < .05$) revealed that preceding word frequency increased the likelihood of hesitations more strongly after shorter words than after longer words ($\beta = 3.47$, $p = .012$); see Table 1 for all model coefficients.

[7] A higher semantic similarity indicates competition between all semantically possible activated words creating interference and subsequent selection difficulty (Levelt et al., 1999; Roelofs, 1992)

3.2 Linguistic predictors of repetitions

The frequency ($p < .05$), integration cost ($p < .05$), bigram probability ($p < .05$), and phonological complexity ($p < .001$) of the following word, as well as the phonological complexity of the preceding word ($p < .001$), significantly predicted the presence of repetitions. Repetitions were more likely to occur when both the preceding and following words were phonologically complex and when the following word had a higher bigram probability. Conversely, higher frequency and integration cost of the following word reduced the likelihood of repetitions. A significant interaction effect between the frequency and semantic similarity of the following words ($p < .01$) revealed that frequency reduced repetitions more strongly in high-similarity contexts than in low-similarity contexts ($\beta = 18.8$, $p = .001$).

3.3 Prosodic predictors of hesitations

At the sentence level, duration significantly predicted hesitation ($p < .05$), indicating that for sentences of the same length, longer durations are associated with a higher likelihood of hesitation. Average pitch ($p = .67$) and average intensity ($p = .64$) of the sentence did not significantly predict hesitations. However, descriptive analysis of time-normalized pitch contours at 10% intervals revealed consistently lower pitch at points where hesitations occurred (see Figure 6 in the Appendix).

We then compared the prosodic properties of words before and after a hesitation, as well as with words corresponding to those positions in fluent sentences. Hesitations significantly increased the duration of the surrounding words ($p < .001$) (see Figure 2). The main effect of word position ($p = .62$) and the hesitation $\times$ position interaction ($p = .10$) were not significant. Hesitation tokens themselves were significantly longer than their matched fluent counterparts (hesitations: M = 301 ms, SE = 10.9; fluent: M = 225 ms, SE = 9.0; $\beta = 0.31$, $p < .001$) (see Figure 7 in the Appendix).

Pitch was analyzed using an *lmer* model with random intercepts for speaker, gender, and sentence length. No significant effects were found for hesitations ($p = .14$) and position ($p = .28$). A significant hesitation $\times$ position interaction was found for pitch ($p = .02$), driven by a decrease in pitch in the word following hesitations, though the pairwise difference was not significant ($p = .082$) (see Fig-

Predictor	Hesitation				Repetition			
	Coef	SE	z	p	Coef	SE	z	p
Preceding word frequency	-2.97	1.16	-2.56	0.0106*	-0.91	0.53	-1.73	0.084
Following word frequency	-0.52	1.00	-0.53	0.600	-1.65	0.69	-2.40	0.0165*
Preceding integration cost	2.95	1.21	2.44	0.0149*	-0.00	0.32	-0.01	0.994
Following integration cost	1.05	0.40	2.65	0.0080**	-0.49	0.25	-1.99	0.0468*
Preceding surprisal	1.94	0.72	2.70	0.0069**	-0.06	0.41	-0.15	0.884
Following surprisal	-0.88	0.70	-1.26	0.208	-0.63	0.48	-1.32	0.188
Preceding semantic similarity	1.65	1.10	1.49	0.135	-0.33	0.44	-0.74	0.462
Following semantic similarity	3.73	1.27	2.93	0.0034**	0.08	0.49	0.16	0.872
Preceding bigram probability	-0.76	0.45	-1.69	0.090	-0.48	0.28	-1.69	0.090
Following bigram probability	-1.29	0.70	-1.84	0.066	1.10	0.43	2.54	0.0112*
Preceding word length	1.92	1.09	1.76	0.079	-	-	-	-
Following word length	4.09	1.24	3.31	0.0010***	-	-	-	-
Preceding phonological complexity	11.87	3.67	3.23	0.0012**	5.65	1.13	4.99	< 0.001***
Following phonological complexity	8.30	2.58	3.22	0.0013**	6.73	1.19	5.68	< 0.001***
Preceding word frequency × word length	-2.53	1.01	-2.52	0.0118*	-	-	-	-
Following word frequency × semantic similarity	-	-	-	-	-2.33	0.73	-3.21	0.0013**

*p<0.05, **p<0.01, **p<0.001

Table 1: Fixed effects from *glmer* models predicting hesitations and repetitions. Shaded rows indicate statistically significant effects.

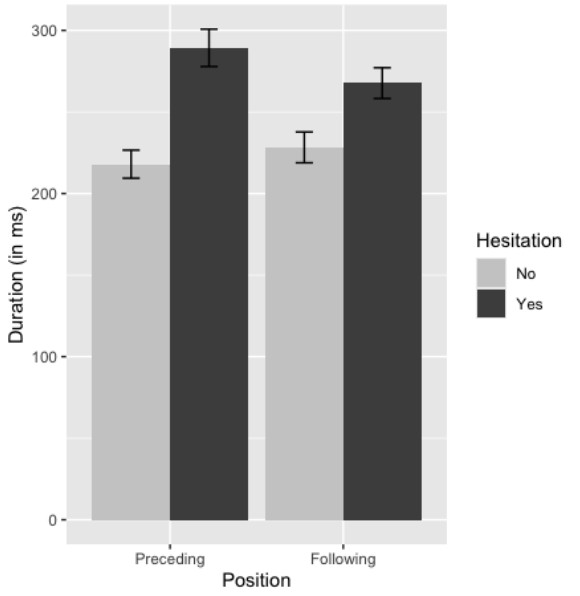

Figure 2: Mean duration of words preceding and following hesitations compared to words in no hesitation contexts.

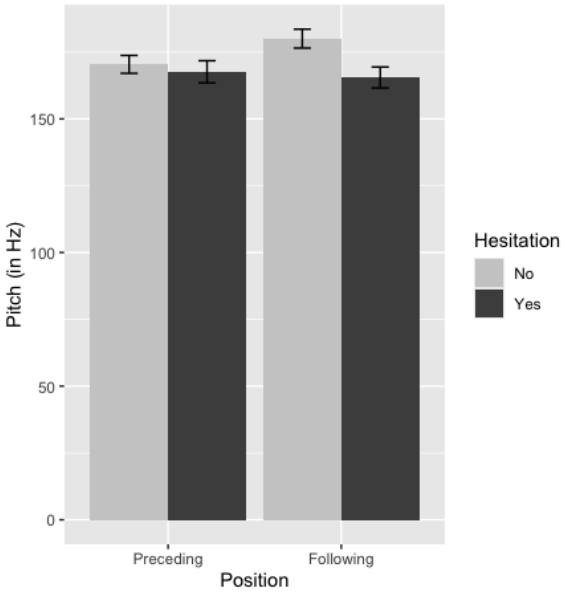

Figure 3: Mean pitch of words preceding and following hesitations compared to words in no hesitation contexts.

ure 3). However, the average pitch of hesitations was significantly lower than that of their matched fluent counterparts (hesitations: M = 162 Hz, SE = 4.28; fluent: M = 175 Hz, SE = 3.66; $\beta = -0.12$, $p = .003$) (see Figure 8 in the Appendix).

The average intensity of words was compared using an *lmer* model, which included random intercepts for speaker, gender, and sentence length. While the presence of hesitations ($p = .10$), word position ($p = .35$), or their interaction ($p = .20$) failed to reach significance (see Figure 4), the intensity of the hesitations themselves was significantly lower than that of their matched fluent counterparts (hesitations: M = 63.7 db, SE = 0.57; fluent: M = 64.9 db, SE = 0.41; $\beta = 0.04$, $p < .01$) (see Figure 9 in the Appendix).

4 Discussion

We investigated the determinants of hesitations and repetitions in spontaneous Hindi speech using the IIT Delhi Hindi Dialog Corpus (Pareek et al., 2025). We demonstrate that both hesitations and repetitions are sensitive to lexical, syntactic, and phonological factors (see Table 1). Our findings are consistent with the previous literature on disfluencies

Prosodic Feature	Fixed Effect	Coef	*SE*	*t*	*p*
Duration	Hesitation	−52.39	12.40	−4.23	< .001**
	Position	4.77	9.60	0.50	.62
	Hesitation × Position	−31.56	19.19	−1.65	.10
Pitch	Hesitation	6.94	4.68	1.48	0.14
	Position	−2.99	2.75	−1.09	0.28
	Hesitation × Position	−12.50	5.50	−2.27	0.02*
Intensity	Hesitation	1.18	0.72	1.64	0.10
	Position	−0.39	0.41	−0.93	0.35
	Hesitation × Position	−1.07	0.83	−1.29	0.20

*p<0.05, **p<0.01, **p<0.001

Table 2: Fixed effects from *lmer* models examining prosodic predictors of hesitation. Shaded rows indicate statistically significant effects.

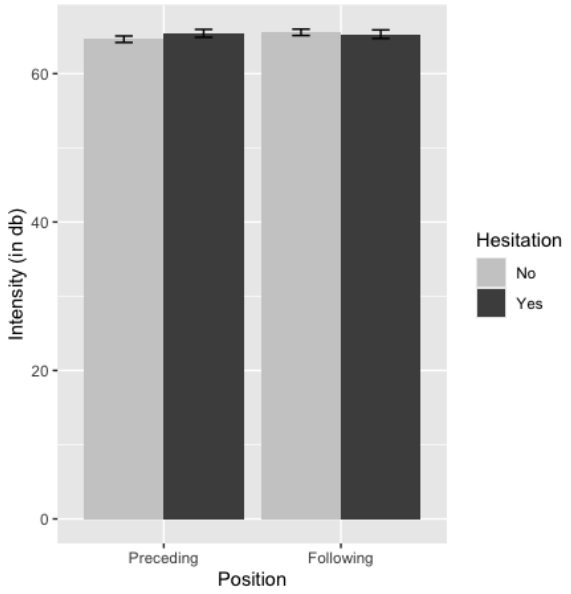

Figure 4: Mean intensity of words preceding and following hesitations compared to words in no hesitation contexts.

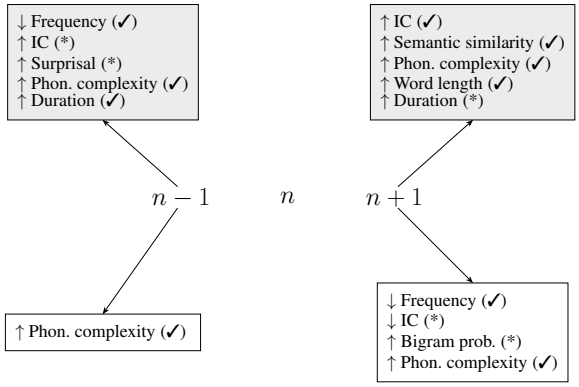

Figure 5: Summary of linguistic and prosodic predictors exhibiting statistically significant effects across position (Preceding $(n-1)$ vs. Following $(n+1)$) and speech error type (Hesitation vs. Repetition). A speech error is denoted by n. Grey panels correspond to hesitation effects, whereas white panels correspond to repetition effects. A check mark (✓) indicates a predicted effect; a star $(*)$ indicates an unexpected result.

for duration and integration cost of the following word (Dammalapati et al., 2021; Shriberg, 1994), as well as prosodic flattening (Zayats and Ostendorf, 2019; Székely et al., 2017). At the same time, our results show that the properties of preceding and following words differentially predict both disfluency types. These findings supplement the existing literature with evidence from Hindi, a typologically distinct head-final language, and reveal several salient patterns across disfluency types that we detail below.

Our first salient finding is that hesitations (as opposed to repetitions) increase when processing load at the previous word is high (see Figure 5). In particular, hesitations increase when the integration cost, surprisal, and phonological complexity of the preceding word increase, and hesitations decrease when the preceding word frequency is high. At the prosodic level, longer durations of surrounding words and notably lower pitch at hesitation sites further support the processing load hypothesis. This set of results is quite interesting because much previous work has found that the role of upcoming difficulty is more critical for hesitations (e.g., Dammalapati et al., 2021). This effect of the preceding element suggests that the processing load in the preceding context could limit the resources available for prior planning, which could reduce the scope of planning. This reduced planning scope manifests as hesitation. Given that such effects of the previous word have not been found in English (e.g., Dammalapati et al., 2021), we speculate that this could be an example of typological distinction during sentence planning. Recent research suggests that intervening dependency heads determine the syntactic complexity in a region (e.g., Yadav et al., 2022). As Hindi is a head-final language and En-

glish is head-initial, planning the resource-intensive heads that come later could affect resource allocation during production, effectively creating downstream difficulty (cf. Scontras et al., 2015; Gibson et al., 2000). In other words, an increase in difficulty encountered at a word would lead to increased resource utilization to process it, which could reduce the available resources needed to plan the upcoming linguistic material (Fox Tree, 1995).

One possible way to test this resource limitation proposal would be to employ forward surprisal (Ranjan et al., 2020), which captures the conditional probability of a word given the *following* context.[8] The idea is that if the processing difficulty affects planning, then the forward surprisal at a word would get attenuated when the preceding word difficulty is high. Results show that the critical interaction between the frequency of the preceding word and the following forward surprisal was not significant. See Table 3 in the Appendix section for more details. This preliminary analysis did not find evidence for reduced planning hypothesis as operationalized by forward surprisal. Future work should investigate this hypothesis more thoroughly by employing other metrics.

As opposed to the influence of the preceding word on hesitations, the following word properties showed a stronger alignment with previous literature (Dammalapati et al., 2021; Arnold et al., 2003), such that integration cost, semantic similarity, word length, and phonological complexity collectively predicted hesitations in the expected direction (see Figure 5).

In contrast to hesitations, for repetitions, the majority of preceding word predictors showed no effect. Critically, the fundamental distinction between hesitations and repetitions emerges in their position-sensitivity. Hesitations are predicted by properties of both preceding and following words, whereas repetitions are predicted primarily by properties of the following word. This asymmetry suggests that hesitations and repetitions arise from different stages of speech production (Fraundorf and Watson, 2013; Lickley, 2017).

Interestingly, phonological complexity was an exception to this asymmetry, predicting both hesitations and repetitions. We speculate that this is because phonological complexity imposes processing costs in two possible ways, depending on the word's position: a complex preceding word sustains demands on the perceptual loop beyond its articulation (Levelt, 1983, 1989), depleting the resources available for planning upcoming material. A complex following word disrupts the incremental attachment of segments to a metrical frame and syllabic gesture computation (Levelt, 1989). Failure at the initial slot results in hesitation, whereas a mid-word disruption forces the speaker to repeat already-articulated material, resulting in a repetition. This effect demonstrates that phonological encoding is an important constraint on speech fluency (Levelt, 1983; Levelt and Cutler, 1983). A bottleneck at the phonological level can notably lead to speech disruptions (Levelt, 1989) manifesting as hesitations and repetitions. This effect is further supported by our prosodic analysis, where we show that hesitations are surrounded by words with longer durations, which may provide additional time for planning upcoming words (Dammalapati et al., 2021; Bell et al., 2009). In contrast, the lower pitch and intensity of hesitation tokens themselves indicate reduced attention to prosodic modulation during articulation when planning load is high, as speakers allocate cognitive resources to manage the disruption (Zayats and Ostendorf, 2019; Székely et al., 2017).

Indeed, for repetitions, only the frequency of the following word and phonological complexity aligned with our predictions (see Section 2.3). Notably, the integration cost of the following word decreased repetitions, whereas a higher bigram probability of the following word increased repetitions, contradicting the prediction that lower predictability triggers disruptions (Hale, 2001; Sen, 2020; Arnold et al., 2007).

As mentioned earlier, although some properties of the preceding words guided the distribution of hesitations and repetitions, the properties of the following word influenced both phenomena. These findings align with models of incremental sentence production (Kempen and Hoenkamp, 1987), in which lexical access, syntactic integration, and phonological encoding and planning impose constraints on sentence processing and, in turn, its fluency (Fraundorf and Watson, 2013; Lickley, 2017).

Overall, our study extends previous work on speech disfluencies by showing that hesitations and repetitions arise from processing difficulties at distinct stages of the production pipeline and are selectively sensitive to different predictors depending on word position (Fraundorf and Watson, 2013; Lickley, 2017). Cross-linguistic evidence from Hindi

[8] Thanks to the anonymous reviewers for this suggestion.

highlights the need to move beyond head-initial languages in disfluency research, as typological organization fundamentally shapes how and where disruptions emerge. Future work should extend this line of inquiry to other head-final and typologically diverse languages to establish the generalizability of these findings.

Conclusion

This work presents a comprehensive investigation of how lexical, syntactic, phonological, and prosodic factors collectively predict disfluencies in Hindi. Our results reveal both similarities as well as dissimilarities between hesitations and repetitions. For example, factors such as phonological complexity robustly predict both these disfluencies. On the other hand, while hesitations are sensitive to cumulative processing pressures across both preceding and following words, repetitions are driven primarily by properties of the following word. These results show that hesitations and repetitions differ in terms of their determinants and locus. Our research highlights how the processes of incrementality/planning and monitoring in an SOV language like Hindi could be different from those in an SVO language like English (cf. Zafar, 2025). Future research should examine the universality of the observed patterns across typologically diverse languages to deepen our understanding of the production process cross-linguistically.

References

Jennifer E. Arnold, Michael Fagnano, and Michael K. Tanenhaus. 2003. Disfluencies signal thee, um, new information. *Journal of Psycholinguistic Research*, 32(1):25–36.

Jennifer E. Arnold, Carla L. Hudson Kam, and Michael K. Tanenhaus. 2007. If you say thee uh-you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):914–930.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve 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 Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111.

Elizabeth R. Blackmer and Janet L. Mitton. 1991. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39(3):173–194.

Kathryn Bock and Willem J. M. Levelt. 1994. Language production: Grammatical encoding. In *Handbook of Psycholinguistics*. Academic Press.

Weenink Boersma. 2012. Boersma, paul; weenink, david. *Praat: Doing phonetics by computer*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Heather Bortfeld, Silvia D. Leon, Judith E. Bloom, Michael F. Schober, and Susan E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147.

Susan E Brennan and Maurice Williams. 1995. The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, 34(3):383–398.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

Alexandra Canavan and George Zipperlen. 1996. CALLFRIEND Hindi. Web download. LDC Catalog No. LDC96S52.

Chiara Celata. 2020. Bottom-up probabilistic information in visual word recognition: interactions with phonological and morphological functions. *Language Sciences*, 78:101267.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Francesca Cossavella and Javier Cevasco. 2021. The importance of studying the role of filled pauses in the construction of a coherent representation of spontaneous spoken discourse. *Journal of Cognitive Psychology*, 33(2):172–186.

Sumanth Dammalapati, Rajakrishnan Rajkumar, Saurabh Ranjan, and Saurabh Agarwal. 2021. Effects of duration, locality, and surprisal in speech disfluency prediction in english spontaneous speech. *Society for Computation in Linguistics*, 4(1):91–101.

Fernanda Ferreira. 1993. Creation of prosody during sentence production. *Psychological Review*, 100(2):233–253.

Jean E. Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34(6):709–738.

Jean E Fox Tree. 2001. Listeners' uses of um and uh in speech comprehension. *Memory & cognition*, 29(2):320–326.

Scott H. Fraundorf and Duane G. Watson. 2013. Alice's adventures in um-derland: Psycholinguistic sources of variation in disfluency production. *Language, Cognition and Neuroscience*, 29(9):1083–1096.

Victoria Fromkin. 1973. *Speech errors as linguistic evidence*, volume 77. Walter de Gruyter.

Victoria A Fromkin. 1971. The non-anomalous nature of anomalous utterances. *Language*, pages 27–52.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Merrill F. Garrett. 1975. The analysis of sentence production. In *Psychology of Learning and Motivation*, volume 9, pages 133–177. Elsevier.

Edward Gibson and 1 others. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Frieda Goldman-Eisler. 1961. A comparative study of two hesitation phenomena. *Language and Speech*, 4(1):18–26.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Charles F Hockett. 1967. Where the tongue slips, there slip i. *To honor Roman Jakobson*, 2:910–936.

Daniel Jurafsky and James H Martin. 2013. *Speech and Language Processing: Pearson New International Edition PDF eBook*. Pearson Higher Ed.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Gerard Kempen and Eduard Hoenkamp. 1987. An incremental procedural grammar for sentence formulation. *Cogn. Sci.*, 11:201–258.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, and 1 others. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Willem J. M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press.

Willem J. M. Levelt and Anne Cutler. 1983. Prosodic marking in speech repair. *Journal of Semantics*, 2(2):205–218.

Willem J. M. Levelt, Ardi Roelofs, and Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–38.

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

Robin Lickley. 2017. Disfluency in typical and stuttered speech. *Book series Studi AISV*, 3:373–387.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Sibout Govert Nooteboom. 1980. Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In *Errors in linguistic performance: Slips of the tongue, ear, pen and hand/ed. By Victoria A. Fromkin*, pages 87–95. Academic Press Inc.

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1915–1921.

Benu Pareek, Mudafia Zafar, Meghna Hooda, Karan Yadav, Ashwini Vaidya, and Samar Husain. 2025. Iit delhi dialogue corpus: a quantitative analysis of a spoken corpus of hindi. *Language Resources and Evaluation*, pages 1–36.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Albert Postma and Herman Kolk. 1993. The covert repair hypothesis: Prearticulatory repair processes in normal and stuttered disfluencies. *Journal of Speech and Hearing Research*, 36(3):472–487.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sidharth Ranjan, Rajakrishnan Rajkumar, Sumeet Agarwal, and IISER Bhopal. 2020. Forward surprisal models production planning in reading aloud. In *Proceedings of the 26th Architectures and Mechanisms for Language Processing Conference (AMLaP), Potsdam, Germany. University of Potsdam*.

Ardi Roelofs. 1992. A spreading-activation theory of lemma retrieval in speaking. *Cognition*, 42(1–3):107–142.

Douglas Roland, Hongoak Yun, Jean-Pierre Koenig, and Gail Mauner. 2012. Semantic similarity, predictability, and models of sentence processing. *Cognition*, 122(3):267–279.

Gregory Scontras, William Badecker, Lisa Shank, Eunice Lim, and Evelina Fedorenko. 2015. Syntactic complexity effects in sentence production. *Cognitive science*, 39(3):559–583.

Priyanka Sen. 2020. Speech disfluencies occur at higher perplexities. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 92–97. Association for Computational Linguistics.

Elizabeth Ellen Shriberg. 1994. Preliminaries to a theory of speech disfluencies. *Doctoral dissertation, University of California at Berkeley*.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Éva Székely, Joseph Mendelson, and Joakim Gustafson. 2017. Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies. In *Proceedings of INTERSPEECH*, pages 804–808.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

Michael S Vitevitch, Jonna Armbrüster, and Shinying Chu. 2004. Sublexical and lexical representations in speech production: effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):514.

Michael S Vitevitch and Paul A Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of memory and language*, 40(3):374–408.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Bodo Winter. 2019. *Statistics for linguists: An introduction using R*. Routledge.

Himanshu Yadav, Shubham Mittal, and Samar Husain. 2022. A reappraisal of dependency length minimization as a linguistic universal. *Open Mind: Discoveries in Cognitive Science*, 6:147–168.

Mudafia Zafar. 2025. Planning scope in production: How grammar shapes incrementality. *Doctoral dissertation, Indian Institute of Technology Delhi*.

Vicky Zayats and Mari Ostendorf. 2019. Giving attention to the unexpected: Using prosody innovations in disfluency detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 86–95, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix

A.1 Linguistic feature computation

The procedure used to compute each linguistic feature is described below.

Word Frequency Raw token counts were tallied across the entire corpus using Python's `Counter`. These counts were converted to the Zipf scale using the formula:

$$\text{Zipf}(w) = \log_{10}\left(\frac{f_w}{N} \times 10^9\right) + 3$$

where f_w is the raw count of the word w and N is the total number of tokens in the corpus (Van Heuven et al., 2014).

Integration Cost Integration cost was computed as the linear distance between a syntactic head and its dependent, extracted directly from the dependency-annotated structures provided in the corpus (Gibson et al., 2000; Futrell et al., 2015).

Surprisal IndicBERT (Kakwani et al., 2020) (`ai4bharat/indic-bert`), a masked language model pre-trained on Hindi and other Indian languages, was used to compute surprisal (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2023; Oh and Schuler, 2023). The context was derived by joining all tokens preceding the target word.[9] The context of the following word did not include the hesitation token, as it may not

[9]We note that with regard to surprisal computation, Wilcox et al. (2023) found that multilingual language models perform similar to monolingual language models. In addition, their work shows that BERT-based masked models perform as well as autoregressive models. Finally, in our study, the context for surprisal calculation was limited to already spoken words, thus ensuring for the psychological validity of the surprisal values (cf. Wilcox et al., 2023).

have a lexical-level representation in the trained model, and its inclusion would confound the probability estimates. The context and target word were passed through the model to extract logits. The softmax function was applied to the logits to obtain the probability of the target token given its context. Surprisal was then computed as:

$$S(w_{n-1}) = -\log P(w_{n-1} \mid w_1, \ldots, w_{n-2})$$

$$S(w_{n+1}) = -\log P(w_{n+1} \mid w_1, \ldots, w_{n-1})$$

Forward Surprisal Forward surprisal was computed based on the same principles as surprisal stated above, the distinction being that the context was limited to unspoken upcoming words (Ranjan et al., 2020).

Semantic Similarity Hindi FastText vectors (`cc.hi.300.bin`) (Bojanowski et al., 2017) were used to extract vector embeddings. The context representation was the mean of the embeddings of all words to the left of the target position. Similar to surprisal, hesitation tokens were excluded from the context of the following word. Cosine similarity was computed between this mean vector and the target word's embedding using `scikit-learn` (Pedregosa et al., 2011).

Bigram Probability Unigram and bigram counts were collected from all sentences in the corpus. Bigram probability was computed as $C(w_{n-2}, w_{n-1})/C(w_{n-2})$. When a bigram was unavailable, for instance, when the hesitation appeared in the second position in the sentence and no two preceding words existed, a unigram fallback $C(w_{n-1})/N$ was applied for the preceding word and $C(w_{n+1})/N$ for the following word, where N is the total number of tokens in the corpus.

Word Length Word length was measured as the number of syllables in the target word, obtained using the `indicnlp` orthographic syllabifier for Hindi (Kunchukuttan et al., 2020).

Phonological Complexity Words were syllabified using the `indicnlp` orthographic syllabifier for Hindi (Kunchukuttan et al., 2020), such that syllables typically corresponded to consonant-vowel (CV) sequences. Phonological complexity was computed as the product of unigram and bigram syllable probabilities across the word. Since the first syllable has no preceding syllable, its probability was estimated as a unigram frequency in the full corpus. For all subsequent syllables, bigram transition probabilities were used, conditioned on the preceding syllable within the corpus.

A.2 Additional plots

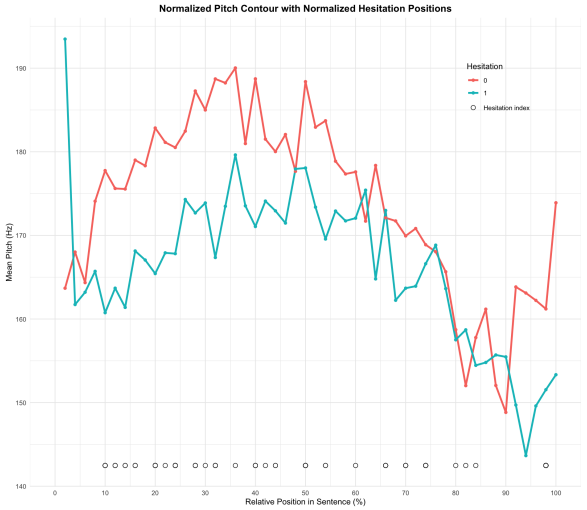

Figure 6: Normalized pitch contours for sentences with (red) and without (blue) hesitations. Hesitation sites show markedly lower pitch.

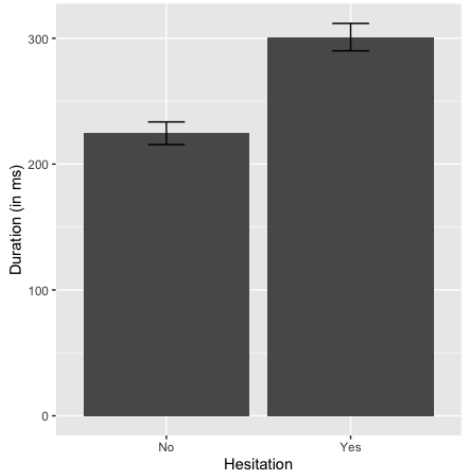

Figure 7: Mean duration of filled pauses and their corresponding no-hesitation tokens

Predictor	Hesitation			
	Coef	SE	z	p
Preceding word frequency	-2.87	1.29	-2.22	0.02603*
Following word frequency	-0.48	1.05	-0.45	0.64976
Preceding integration cost	2.80	1.13	2.48	0.01320*
Following integration cost	0.98	0.39	2.49	0.01281*
Preceding surprisal	1.97	0.79	2.47	0.01339*
Following surprisal	-6.46	3.32	-1.94	0.05182.
Following forward surprisal	5.91	3.13	1.88	0.05942.
Preceding semantic similarity	1.61	1.17	1.38	0.16791
Following semantic similarity	3.53	1.40	2.51	0.01207*
Preceding bigram probability	-0.62	0.45	-1.37	0.16981
Following bigram probability	-1.29	0.75	-1.71	0.08716.
Preceding word length	1.70	1.15	1.48	0.13948
Following word length	4.53	1.55	3.12	0.00181**
Preceding phonological complexity	11.60	3.96	2.93	0.00337**
Following phonological complexity	9.11	2.99	3.04	0.00236**
Preceding word frequency × word length	-2.52	1.10	-2.27	0.02280*
Preceding word frequency × Following forward surprisal	-0.25	0.58	-0.42	0.67060

*p<0.05, **p<0.01, **p<0.001

Table 3: Fixed effects from *glmer* models predicting hesitations. Shaded rows indicate statistically significant effects.

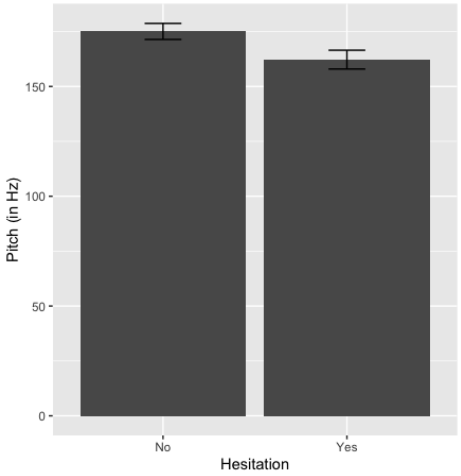

Figure 8: Mean pitch of filled pauses and their corresponding no-hesitation tokens

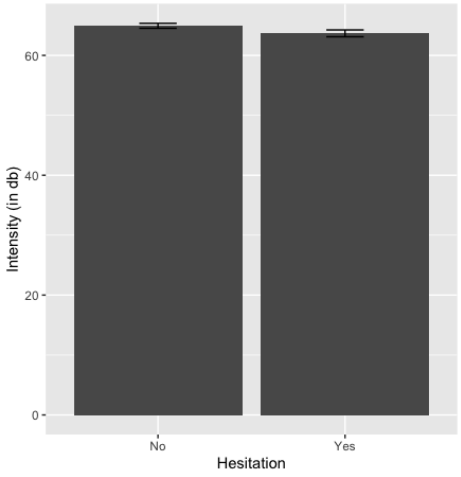

Figure 9: Mean intensity of filled pauses and their corresponding no-hesitation tokens