A Systematic Exploration of Linguistic Phenomena in Spoken Hindi: Resource Creation and Hypothesis Testing

Aadya Ranjan
IIIT Hyderabad
aadya.ranjan@research.iiit.ac.in

Sidharth Ranjan
University of Stuttgart
sidharth.ranjan03@gmail.com

Rajakrishnan Rajkumar
IIIT Hyderabad
raja@iiit.ac.in

Abstract

This paper presents a meticulous and well-structured approach to annotating a corpus of Hindi spoken data. We deployed 4 annotators to augment the spoken section of the EMILLE Hindi corpus by marking the various linguistic phenomena observed in spoken data. Then we analyzed various phonological (sound deletion), morphological (code-mixing and reduplication) and syntactic phenomena (case markers and ambiguity), not attested in written data. Code mixing and switching constitute the majority of the phenomena we annotated, followed by orthographic errors related to symbols in the Devanagiri script. In terms of divergences from written form of Hindi, case marker usage, missing auxiliary verbs and agreement patterns are markedly distinct for spoken Hindi. The annotators also assigned a quality rating to each sentence in the corpus. Our analysis of the quality ratings revealed that most of the sentences in the spoken data corpus are of moderate to high quality. Female speakers produced a greater percentage of high quality sentences compared to their male counterparts. While previous efforts in corpus annotation have been largely focused on creating resources for engineering applications, we illustrate the utility of our dataset for scientific hypothesis testing. Inspired from the Surprisal Theory of language comprehension (Hale, 2001; Levy, 2008), we validate the hypothesis that sentences with high values of lexical surprisal are rated low in terms of quality by native speakers, even when controlling for sentence length and word frequencies in a sentence.

1 Introduction

The availability of high-quality linguistic resources plays a pivotal role in the field of computational linguistics from both theoretical and application-oriented perspectives. Hindi (Indo-Aryan language, Indo-European language family) is considered a medium-resource language primarily spoken in the Indian subcontinent. In the context of spoken language resources, the following datasets (inter-alia) can be considered as pioneering efforts in corpus creation: Simulated emotion Hindi speech corpus (Koolagudi et al., 2011), Indic speech database (Prahallad et al., 2012), LDC-IL Hindi raw speech corpus (Choudhary and Rao, 2020), and Hindi-Urdu Treebank corpus (Bhatt et al., 2009, HUTB), consisting of both written and spoken data. However, except for HUTB, the majority of dataset developments in the Indian context have predominantly focused on engineering applications with relatively less emphasis placed on scientific hypothesis testing and language processing research. In this work, we present our preliminary efforts to annotate the Hindi spoken section within the publicly available EMILLE corpus (McEnery et al., 2000), with information pertaining to various linguistic phenomena. The key objectives of our corpus annotation project are summarized below:

1. **Resource creation**: We deployed 4 annotators to augment the spoken section of the EMILLE Hindi corpus by marking various linguistic phenomena, *viz.*, phonological (sound deletion), morphological (code-mixing and reduplication) and syntactic phenomena (case markers and ambiguity). Then we analyzed those phenomena which are not attested frequently in written data.

2. **Hypothesis testing**: We test the hypothesis that sentences with high values of lexical surprisal are rated low in terms of quality by native speakers, even when controlling for sentence length and word frequencies. Our hypothesis is motivated by the Surprisal Theory (Hale, 2001; Levy, 2008), an information-theoretic characterization of language comprehension.

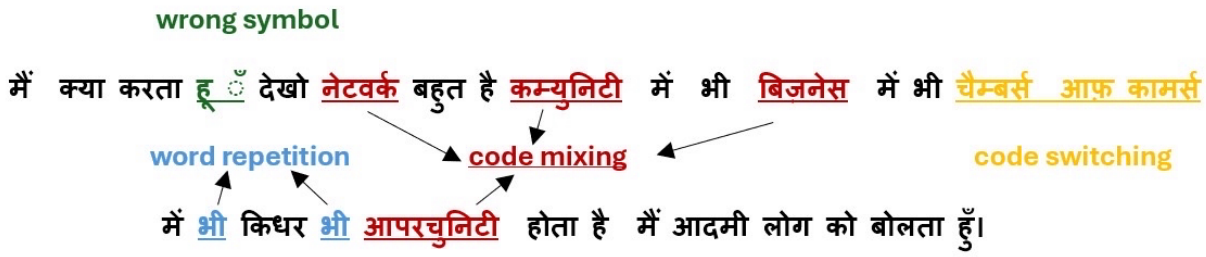

Figure 1: Spoken Hindi sentence taken from the EMILLE corpus

To achieve these objectives, we extensively cleaned the Hindi spoken section of the EMILLE corpus, removing inconsistencies and errors not representative of the language. Subsequently, we deployed 4 annotators to augment the corpus with annotations encoding various linguistic phenomena commonly found in speech, as illustrated in Figure 1. Code mixing and switching constitute the majority of the phenomena we annotated, followed by orthographic errors related to symbols in the Devanagiri script. In terms of divergences from written form of Hindi, case marker usage, missing auxilliary verbs and agreement patterns are markedly distinct for spoken Hindi. Our annotators also assigned a quality rating to original sentences based on their grammaticality, coherence, clarity, and overall effectiveness in communication. Our analysis of the quality ratings revealed that the most of the sentences in the spoken data corpus are of moderate to high quality. Female speakers produced a greater percentage of high quality sentences compared to their male counterparts, due to the tendency to adhere to the prescriptive norms of the language. While previous efforts in corpus annotation have been largely focused on creating resources for engineering applications, we illustrate the utility of our dataset for scientific hypothesis testing.

Behavioural experiments and corpus analyses are two most prominent methods in psycholinguistics (Traxler and Gernsbacher, 2011). The corpus data contains naturally occurring sentences and thus offers an ecologically valid paradigm to test cognitively motivated hypothesis (Gries, 2005; Rajkumar et al., 2016; Ranjan et al., 2022b). It can also complement outcomes from behavioral methods that use carefully controlled stimuli designed by experimenters (Demberg and Keller, 2008; Ranjan et al., 2022a). In this work, leveraging our spoken Hindi dataset, we validate the aforementioned hypothesis (motivated by Surprisal Theory) using a Linear Mixed Model (LMM, Pinheiro and Bates, 2000) to predict the sentence quality rating of the sentences in the dataset we created (annotators and gender of the speakers serving as the GLM intercept terms). Surprisal Theory defines the *surprisal* of the $(k+1)^{th}$ word, w_{k+1}, as the negative logarithm of conditional probability of word, w_{k+1} given the preceding context, which can be either sequence of words or a syntactic tree. Both these kinds of surprisal have been shown to predict eye movement durations in language processing (Demberg and Keller, 2008; Smith and Levy, 2013). Frequency-based controls are based on long-standing findings from the literature attesting that high frequency words are processed faster than their low frequency counterparts on account of higher activation resulting from increased exposure (Morton, 1969; Forster and Chambers, 1973). We showed that high values of lexical surprisal predict lower quality ratings, corroborating prior findings in the Hindi corpus-based sentence processing literature (Ranjan et al., 2022c; Ranjan and van Schijndel, 2024).

Our primary contribution is the development of a Hindi linguistic resource, created by augmenting the EMILLE corpus of spoken Hindi with linguistically motivated annotations and quality ratings. We believe this work will facilitate both engineering applications and scientific research aimed at validating and advancing theories of language production and comprehension.

The paper is structured as follows: Section 2 describes our methods, Section 3 presents our experiments and Section 4 summarizes the implications of our main findings and outlines a plan for future research. The Appendix provides the corpus annotation manual used by our annotators to create the dataset described in our work.

2 Methodology

This section describes the methodology of corpus annotation adopted in this work. Corpus annotation involves enriching a corpus with linguistic and other information through manual or automatic methods, serving theoretical or practical purposes (Gries and Wulff, 2009). The following subsections describe the actual procedure we adopted to identify linguistic phenomena of interest and assign quality ratings to each sentence in the spoken Hindi section of the EMILLE corpus (McEnery et al., 2000). Additionally, each file contains a comprehensive header detailing the text's provenance (for example **hin-w-ranchi-news-01-03-22.txt**). The dataset comprises of 19793 sentences transcriptions from spoken conversations on BBC Radio, featuring **168 speakers** (48 females, 120 males) from the following domains:

1. **Read-aloud speech**: News bulletins as well as messages from radio listeners which were read and acknowledged.

2. **Conversational speech**: Conversation between an anchor and invited guests regarding current affairs and entertainment.

In contrast, the EMILLE corpus of written data draws from newspapers like India Info, Ranchi Express, and Web Duniya. A broad-brush comparison between EMILLE spoken and written data revealed that EMILLE written data consists of longer sentences (average sentence length of 44.24 words) compared to the spoken sentences (average sentence length: 33.39 words). Written data exhibits a higher token-type ratio (0.4402) than spoken data (0.2345), indicating richer vocabulary in written language. More systematic comparisons of spoken and written data need to be undertaken to provide a comprehensive picture of the similarities and differences between these modalities.

2.1 Corpus Annotation

Initially, we went through random samples of sentences from the spoken corpus under study and prepared an annotation manual (see Appendix) documenting various linguistic phenomena that are potentially interesting from language production research. Subsequently, 4 different native speakers of the Hindi language were trained to examine sentences and identify various linguistic phenomena mentioned in the annotation manual and provide a quality rating for each of the 19,793 sentences.[1] Table 1 illustrates the annotation process with linguistic examples. The final dataset consisted of the following 5 attributes:

1. **Sentence ID**: To each sentence in the EMILLE corpus, we assigned a unique identifier (ID) in the format (`File_S_Y_Z`) encoding the document number (File), speaker tag (S), paragraph number (Y) and sentence number (Z). For example, the sentence ID: `File1_HF001_1_1` corresponds to the first sentence in paragraph 1 articulated by speaker HF001, and electronically transcribed in text form in the document File1.

2. **Quality rating**: Annotators were asked to rate sentences based on their own understanding or intuitive sense as native speakers. They provided a quality rating (1-ungrammatical; 2-poor; 3-fair; 4-good; 5-excellent) for each sentence along with various linguistic annotations. This subjective assessment helps gauge the linguistic fluency, correctness, coherence, clarity, and overall effectiveness in communication.

3. **Remarks**: Annotators are required to provide detailed explanations regarding what specifically is wrong with the sentence or how it deviates from the established grammatical norms of written Hindi. This qualitative feedback offers valuable insights into the nuances of sentence construction, aiding in pinpointing areas that require improvement.

4. **Raw Sentence**: This denotes the original EMILLE corpus sentence exactly as it appears, without any alterations or corrections. This untouched rendition preserves the authenticity of the original data, enabling accurate analysis and comparison.

5. **Annotated Sentence**: Annotators were asked to correct grammatical errors present in the sentence while ensuring that the meaning and essence of the sentence remain intact. This delicate task of rectifying linguistic inaccuracies while preserving semantic coher-

[1] A sample file containing 100 sentences from our corpus is available via: https://github.com/Aadya38/IIITH_Hindi_Spoken_Corpus

Sentence ID	Quality	Remarks	Sentence
File1_HF001_1_1	4	Code mixing	*do baj kar chaar* ***minut*** *hue hai*
File1_HF001_1_2	3	Missing aux verb	*jab humare pension ki baari* ***(missing hai)*** *toh ...*
File1_HF001_1_3	2	Ambiguous	*agla nagma surekha ji* ***apke nazar*** *karna chahte hai*
File1_HM073_18_1	5	Reduplication	***bahut-bahut*** *shukriya*

Table 1: Corpus annotation examples

Phenomenon	#Occurrences	Phenomenon	#Occurrences
Code mixing	6980	Missing auxiliary verb	526
Symbol error	1703	Agreement error	482
Reduplication	1651	Code switching	452
Word repetition	1123	Case marking error	389
Abrupt ending	543	Ambiguity	333

Table 2: Frequency of linguistic phenomena annotated

ence requires a deep understanding of the language's intricacies.

From a psycholinguistic perspective, sentence ratings reflect the coherence, clarity and comprehensibility of a sentence. The process of assigning sentence quality relies on the intuitions of native speakers. So analyzing these ratings offer a window into the cognitive factors influencing language processing. Acceptability judgments are widely used to validate syntactic theories (Sprouse and Almeida, 2017). We now elaborate on various linguistic annotations employed in this work and elaborate on them further with the help of a quantitative analysis, described in the next section.

3 Experiments

This section provides an analysis of the quality ratings and a summary of various linguistic phenomena pertaining to phonology, morphology, and syntax that were observed and annotated by four native Hindi speakers following our annotation manual. Each linguistic annotation, along with their frequency, is presented in Table 2. Code mixing constitutes the most frequent class of annotations assigned by the raters, followed by orthographic errors and reduplication. Enriching corpora with linguistic information serves a crucial function in developing and testing linguistic theories, in addition to the training machine learning algorithms for engineering applications. So we considered linguistic phenomena pertaining to all aspects of language, *viz.*, phonology, morphology, syntax, semantics, and pragmatics.

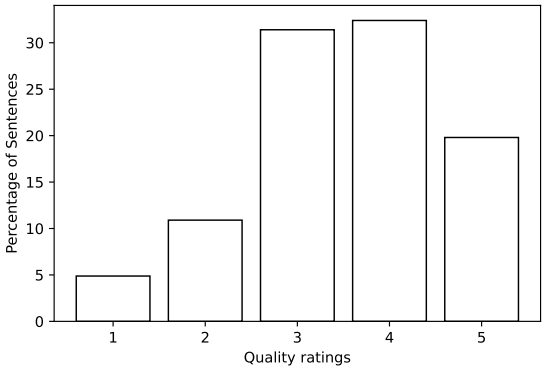

Figure 2: Distribution of quality ratings in the annotated corpus of transcribed speech

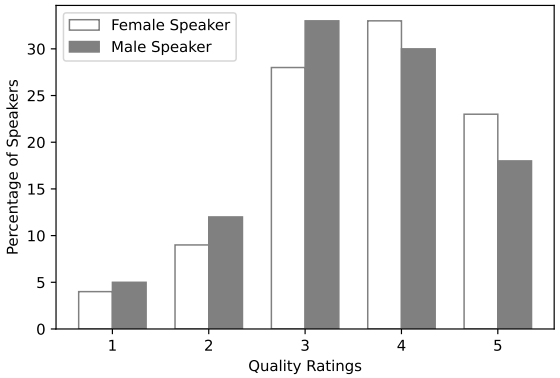

Figure 3: Gender-wise distribution of quality ratings

3.1 Analysis of Quality Ratings

Quality ratings are essential for distinguishing between written and spoken language modes. Despite spoken sentences being perceived as correct by speakers, annotations from native speakers on written sentences often reveal differences. Our corpus stands out from other Hindi written corpora and includes a range of sentence quality from poor to excellent (1-ungrammatical; 2-poor; 3-fair; 4-good; 5-excellent), maintaining authenticity as a true spoken corpus. The mean quality rating (1-5) across entire sentences in the corpus was found to be **3.568**. Thus, the sentences in the corpus are moderately good in terms of quality.

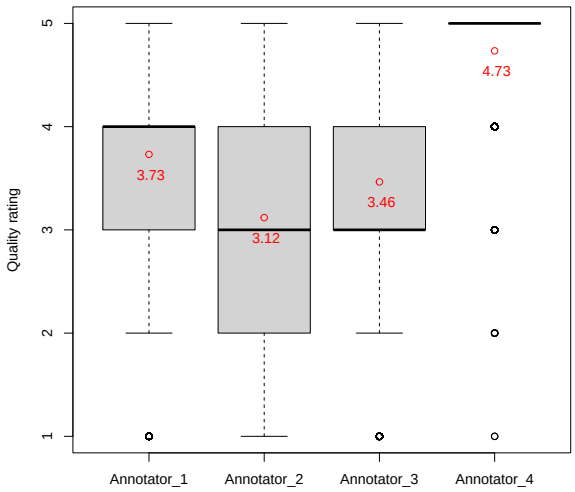

Figure 4: Annotator-wise means of quality ratings

Figure 2 illustrates the distribution of quality ratings in the EMILLE corpus after annotation. Around 60% of the sentences received a rating of 3 (fair) or 4 (good), followed by around 25% sentences rated 5 (excellent). A small percentage of the sentences turned out to be of poor quality. Subsequently, we examined the annotator-wise means of the ratings data. Figure 4 shows the results, indicating that the sentences assigned to Annotator_4 were given a higher mean rating compared to all the other three annotators. There are 2 competing explanations for this finding. The first possibility that one annotator had a more liberal view of sentence quality compared to the others and hence ended up giving a higher score to most sentences. Alternatively, the sentences were really of high quality in that set. Deploying more annotator and computing inter-annotator agreement is the solution to this conundrum, that we plan to explore as a part of future work.

We then explored the role played by the gender of speakers and the perception of sentence quality. This tangent was motivated by the variation induced by gender in terms of speech styles, adherence to prescriptive norms and lexical choice (use of pronouns and honorifics for example). Figure 3 illustrates the gender-specific patterns in the quality ratings. The figure clearly indicates a notable difference in quality ratings between female and male speakers. Female speakers exhibit a higher quality level, with 23% of sentences receiving a 5 quality rating, while male speakers have a slightly lower percentage at 18%. We attribute this finding to the greater adherence of women towards the prescriptive norms of the language. Future work along these lines would reveal more interesting aspects about gender norms and stereotypes in communication.

3.2 Phonological Analysis

Our corpus annotation efforts revealed the following types of phonological phenomena:

Palatal Fronting: In this phonetic shift, the fricative consonants '/sh/' and '/z/' are replaced by sounds produced further forward on the palate, closer to the front teeth (Francisco and Wertzner, 2017). '/sh/' is substituted with '/s/', and '/z/' is replaced by /j/ (Keating, 1993), as indicated by example: *dheere dheere* ***saam*** *bhi ho jayegi (it will slowly become evening)*

S-Retraction: This articulatory change involves producing the /sh/ sound with the tongue positioned slightly farther back in the mouth compared to the /s/ sound (Keating, 1993) in examples like *sharhad* vs *sarhad (border)*

Consonant and Vowel Deletion: Consonant deletion occurs when a consonant is omitted at the beginning or end of a syllable (Elbert and Mcreynolds, 1985). Vowel deletion refers to the elimination of an unstressed vowel or the transformation of a diphthong into a monothong by the removal of one of the vowels, can also be known as Schwa deletion (Magdum et al., 2019). The natural pace of speech, faster than that of writing, leads to more frequent omissions of sounds during spoken communication (Dell and Reich, 1981), examples like *khairiyat* vs *kheriyat*.

Symbol error: This refers to incorrect orthography in transcribed words. Hindi follows the Devanagari alphasyllabary-based writing system. However, a common issue we observed was an incorrect placement of the chandrabindu (U + 0981 chandrabindu bengali sign) among others (see Figure 1). Annotators corrected these to conform to the standard conventions of the Devanagiri script (Templin, 2013). These corrections serve crucial purposes in training NLP tools such as taggers and parsers, as consistent training and testing data are essential for machine learning algorithms.

3.3 Morphological Analysis

A summary of the morphological phenomena annotated in our corpus is provided below, along with the help of the following examples (sentence IDs in parentheses):

(1) a. Reduplication (HU900_100_1)

bahut bahut bahut dhanyavad
very very very thanks

Thank you very much.

b. Code mixing (HM003_4_4)

skatish sekreteri Helen Liḍl=ne
Scottish secretary Helen Liddell=ERG
gavarnmeṇṭ choṛ dī hai
government leave give=PFV.F.SG be=PRS

Scottish Secretary, Helen Liddell, has resigned from the government.

c. Word repetition (HM003_40_1)

andaz=se kah rahe hain Navindar
guess from saying be=PRS.PL Navindar
jī **andaz=se**
honorific guess from

Navindar is saying that it based on a guess

Reduplication: Refers to the repetition of the root, stem, or entire word, either exactly or with slight variations in form (Singh, 2005). Repetition often serves to emphasize the significance of a word in speech as in example 1(a).

Code mixing: Refers to the embedding of linguistic units such as words and morphemes of one language into an utterance of another language. We noticed the widespread use of English words like *government* and *secretary* rather than their vernacular equivalents. Disfluencies caused by code-switching and mixing are common in speech and indicate processing constraints (Kim, 2006), example 1(b).

Word repetition: Refers to instances where a word is repeated multiple times within a sentence, potentially indicating a speech error. Repetition may occur due to ongoing cognitive processes, such as a momentary pause to gather thoughts before continuing with another sentence or to emphasize a particular information, example 1(c).

3.4 Syntactic Analysis

A summary of syntactic phenomena annotated in our corpus is discussed below with the help of the following examples (sentence IDs in parentheses):

(2) a. Wrong Agreement (UF003_45_37)

is=ki alava mausam=mein khasi
this=GEN besides weather=LOC intense
sardi bhi pai jaegi
cold also found go=FUT.F.SG

Besides this, intense cold will also be found in the weather.

b. Addition of case (HF001_16_1)

accha=**ki** bat kar=te hain
good GEN thing do-PRS.PTCP.PL
to vah kiya hai
be=AUX so that done be=PRS

When we talk about good things, what has been done?

c. Wrong case (HF001_1_4)

ummeed hai vikend=ke bad meri
hope be-PRS weekend GEN after
awaz=**mein** aap=ko khairiyat=mein
my voice=LOC you=DAT
paya hoga
well-being=LOC found be=FUT

I hope after the weekend, you found me in good health through my voice.

d. Missing case (HF017_26_1)

jaise mai=ne aap=se kaha aap aaksfaṛḍ=ke
as I=ERG you=DAT said you Oxford
nambar to aap=ko de sakti
GEN number then you=DAT give
hun
can=F.SG am

As I said, I can give you the number for Oxford.

e. Missing aux (HM002_349_2)

is=ke alava bhi kya aap=ne
this=GEN besides also what you=ERG
kisi=se sikhi mousiki
someone=ABL learned music

Besides this, did you also learn music from someone?

f. Abrupt ending: HM003_67_2

saudi rajdhani Riyaḍ=mein ek briṭish admi
Saudi capital Riyadh=LOC one British man
goli=se halaaq...
bullet=INST killed

A British man was killed by a bullet in the Saudi capital, Riyadh...

g. Speech Error - HM124_57_4

yani kai bar aap=ke **par** nahin
meaning many times you=GEN on not
lage hote hain
attached be=PST.PTCP be=PRS.PL

Many times, it is not that things are on your side.

Agreement errors: The subject or object of a sentence must align in number and gender with the verb in a sentence. The inaccuracies associated with agreements are fairly evident in spoken Hindi (Comrie, 1984), as shown in Example 2(a).

Ambiguity: Terms are frequently used to convey multiple meanings, creating confusion

for machine comprehension. Certain ambiguous statements exhibit varied syntactic structures due to violations of the binding theory, posing complexities for artificial intelligence in sentence understanding.

Case marking: The case of a word determines its grammatical role as a subject (nominative), direct object (accusative), indirect object, object of a preposition, or possessive form (genitive) (Spencer, 2005). Errors such as adding, or omitting cases are prevalent in spoken language, leading to varied interpretations as exemplified in sentences 2(b), 2(c) and 2(d) below.

Missing auxiliary verb: Modal verbs are vital for indicating verb tenses and expressing likelihood, ability, permission, and obligation. They are occasionally omitted or added unintentionally in the corpus. Auxiliary verb omission is a frequency phenomenon in Hindi (as shown in Table 2 and Example 2(e)).

Abrupt ending: Effective communication requires clarity and coherence. However, in spoken data, speakers frequently leave sentences unfinished, quite distinct from formal written communication, see example 2(f).

Speech Error: Different types of speech errors offer insights into the functioning of different components of the production system. For instance, semantic substitution errors likely reflect the conceptual preparation or lexical selection component of the speech production process (Dell and Reich, 1981), see example 2(g).

Written data showed a higher percentage (16.15%) of case markers compared to spoken data (11.95%), revealing grammatical differences between the two modalities. Additionally, percentage of conjunction and contrastive clauses were found higher in written data compared to the spoken data, aligning with the findings reported for English (Redeker, 1984).

3.5 Hypothesis Testing

In this section, we test the hypothesis that an increase in sentence-level lexical surprisal leads to lower values of quality rating in sentences when controlling for sentence length and sum total of frequencies of words in a sentence. To this end, we trained the following Linear Mixed Model (Pinheiro and Bates, 2000, LMM) to predict the sentence quality rating of the sentences in our dataset:

Predictors	Estimate	Std. Error	t-value
(Intercept)	3.752	0.339	11.069
Log Frequency	0.130	0.008	-2.896
Sentence length	0.334	0.046	7.258
Lexical surprisal	-0.383	0.046	-8.366

Table 3: Fixed effects of an LMM predicting sentence quality rating (19793 data points; all predictors are significant at |t|=2 threshold)

$$Rating \sim Logfrequency + Lexical\ surprisal + Word\ length + (1|Annotator) + (1|Gender)$$

The $lme4$ package in R was used to perform our regression experiments using a very basic model, given below in R GLM format (independent variable $\sim$ dependent variables + 1| random intercept terms to model random effects pertaining to speakers and items). All the independent variables described below were normalized to z-scores:

- SENTENCE LENGTH: Total number of words in a sentence.
- FREQUENCY: Sum total of frequencies of individual words in a sentence. Word frequency, the count of each target word, was obtained from the written section of the EMILLE Hindi corpus.
- LEXICAL SURPRISAL: Trigram surprisal is defined as the negative log of the probability of target word w_{k+1} given two preceding words: $S_{k+1} = -\log P(w_{k+1}|w_{k-1}, w_k)$. For each word in a sentence, we computed this measure using a trigram language model trained on the EMILLE corpus of written text with mixed genre using the SRILM toolkit (Stolcke, 2002) with Good-Turing smoothing.

Subsequently, sentence-level lexical surprisal and frequency were computed by summing the respective per-word values of these measures. We plugged in the logarithm of the frequency sum into the GLM to make it compatible with the surprisal term (in the log-scale by definition). Our results are depicted in Table 3. Our regression

results successfully validate our hypothesis. We found that lexical surprisal is a significant predictor of sentence quality ratings. The negative regression coefficient associated with surprisal indicates that sentences with higher lexical surprisal are associated with lower values of quality ratings compared to their lower surprisal counterparts. Sentences with longer lengths and more frequent words tend to reflect higher quality, as inferred from the positive coefficients of these features. Our experiments involving mean sentence frequency and mean lexical surprisal by dividing their raw values by sentence length also resulted in similar trends.

4 Discussion and Conclusions

We present a Hindi speech corpus created by augmenting the EMILLE spoken Hindi corpus using linguistically motivated annotations and sentence quality ratings obtained by deploying 4 annotators. Code mixing and switching constitute the majority of the phenomena we annotated, followed by orthographic errors related to symbols in the Devanagiri script. In terms of divergences from written form of Hindi, case marker usage, missing auxiliary verbs and agreement patterns are markedly distinct for spoken Hindi. Our analysis of the quality ratings assigned by the annotators revealed that most of the sentences in the spoken data corpus are of moderate to high quality. Female speakers produced a higher percentage of high-quality sentences compared to their male counterparts. Our conjectures is that this is possibly due to their tendency to adhere to the prescriptive norms of the language very closely.

Our augmented Hindi spoken corpus can be used for both engineering applications as well as for scientific hypothesis testing, in contrast to existing corpora which are exclusively oriented towards engineering applications. Using this dataset, we validated our hypothesis that lexical surprisal is a significant predictor of quality ratings by humans. Sentences with higher lexical surprisal are associated with lower values of quality ratings compared to their lower surprisal counterparts. Sentences with longer lengths and more frequent words tend to reflect higher quality. Our findings directly align with prior corpus-based sentence literature, which attests to the correlation between lexical surprisal estimates and human preference judgments in word order choices (Ranjan et al., 2022c; Ranjan and van Schijndel, 2024). Lexical predictability effects in silent reading are known to occur at the initial stages of word processing, where readers activate a set of plausible upcoming words in the given context. Predictability effects are also associated with the early prelexical stage in the lexical access process, pertaining to the visual properties of the script (Staub, 2015).

A key limitation of our work is that we work on transcribed speech, and thus miss out on the acoustic features associated with spoken data. In future work, we plan to overcome this lacuna by creating a Hindi corpus of transcribed speech data using the pipeline developed by Mirishkar et al. (2023). We intend to create 500 hours of speech with annotations of linguistic phenomena discussed in this work as well as disfluency annotation for two levels of annotation for each recording: 1. *Transcript level*: Marking disfluencies in the transcript obtained from automatic speech recognition for audio recording 2. *Signal level*: The start and end times of each disfluency annotated at the transcript level will be noted. Annotators will be employed to manually annotate all the audio files for 5 types of disfluencies such as *filled pauses, prolongation, part-word repetition, word repetition and phrase repetition* and assign a confidence score to the annotation.

In future research, we plan to explore the impact of each type of linguistic phenomenon (*viz.*, phonological, morphological, and syntactic discussed in the paper) on human preference ratings. We also plan to employ more annotators, compute the inter-annotator agreement, and develop a platinum-standard resource for both engineering applications and scientific inquiries. We also envisage a systematic comparison of text and speech modalities of language use in terms of their cognitive properties.

Acknowledgements

We thank the anonymous reviewers of SAFAL-2023 and ICON-2024 conferences for their invaluable comments and feedback. We are grateful to Sharvari Thorat, Saundarya Prakash, and Manushi Dhar, for their contributions as annotators. Finally, we acknowledge financial support from the Department of Science and Technology of India (project no. DST/CSRI/2018/263) and a fellowship from IHub-Data, IIIT Hyderabad.

References

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for Hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 186–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Narayan Choudhary and DG Rao. 2020. The ldc-il speech corpora. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 28–32. IEEE.

Bernard Comrie. 1984. Reflections on verb agreement in hindi and related languages. *Linguistics*.

Gary S Dell and Peter A Reich. 1981. Stages in sentence production: An analysis of speech error data. *Journal of verbal learning and verbal behavior*, 20(6):611–629.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Mary Elbert and Leija V Mcreynolds. 1985. The generalization hypothesis: Final consonant deletion. *Language and Speech*, 28(3):281–294.

Kenneth I. Forster and Susan M. Chambers. 1973. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12(6):627–635.

Danira Tavares Francisco and Haydee Fiszbein Wertzner. 2017. Differences between the production of [s] and [sh] in the speech of adults, typically developing children, and children with speech sound disorders: An ultrasound study. *Clinical linguistics & phonetics*, 31(5):375–390.

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

Stefan Th Gries and Stefanie Wulff. 2009. Psycholinguistic and corpus-linguistic evidence for l2 constructions. *Annual Review of Cognitive Linguistics*, 7(1):163–186.

John Hale. 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 on Language technologies*, NAACL '01, pages 1–8, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

Patricia Keating. 1993. Phonetic representation of palatalization versus fronting. *UCLA Working papers in phonetics*, 85:6–21.

Eunhee Kim. 2006. Reasons and motivations for code-mixing and code-switching. *Issues in EFL*, 4(1):43–61.

Shashidhar G Koolagudi, Ramu Reddy, Jainath Yadav, and K Sreenivasa Rao. 2011. Iitkgp-sehsc: Hindi speech corpus for emotion analysis. In *2011 International conference on devices and communications (ICDeCom)*, pages 1–5. IEEE.

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

Damodar Magdum, T Patil, and M Suman. 2019. Schwa deletion in hindi language speech synthesis. *International Journal of Engineering and Advanced Technology*, 8(6S):211–214.

Anthony McEnery, Paul Baker, Robert Gaizauskas, and Hamish Cunningham. 2000. Emille: Building a corpus of south asian languages. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the new Millennium: MT 2000*.

Ganesh S. Mirishkar, Vishnu Vidyadhara Raju V, Meher Dinesh Naroju, Sudhamay Maity, Prakash Yalla, and Anil Kumar Vuppala. 2023. Iiith-cstd corpus: Crowdsourced strategies for the collection of a large-scale telugu speech corpus. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(7).

John Morton. 1969. Interaction of information in word recognition. *Psychological review*, 76(2):165.

José C Pinheiro and Douglas M Bates. 2000. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56.

Kishore Prahallad, E Naresh Kumar, Venkatesh Keri, S Rajendran, and Alan W Black. 2012. The iiit-h indic speech databases. In *Thirteenth annual conference of the international speech communication association*.

Rajakrishnan Rajkumar, Marten van Schijndel, Michael White, and William Schuler. 2016. Investigating locality effects and surprisal in written english syntactic choice phenomena. *Cognition*, 155:204–232.

Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2022a. Linguistic Complexity and Planning Effects on Word Duration in Hindi Read Aloud Speech. In *In Proceedings of the Society for Computation in Linguistics (SCiL)*, volume 5, page 11.

Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2022b. Locality and expectation effects in Hindi preverbal constituent ordering. *Cognition*, 223:104959.

Sidharth Ranjan and Marten van Schijndel. 2024. Does Dependency Locality Predict Non-canonical Word Order in Hindi? In *Proceedings of the 46th Annual*

Meeting of the Cognitive Science Society, Rotterdam, Netherlands. Cognitive Science Society, Cognitive Science Society.

Sidharth Ranjan, Marten van Schijndel, Sumeet Agarwal, and Rajakrishnan Rajkumar. 2022c. Discourse Context Predictability Effects in Hindi Word Order. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10390–10406, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gisela Redeker. 1984. On differences between spoken and written language. *Discourse processes*, 7(1):43–55.

Rajendra Singh. 2005. *Reduplication in Modern Hindi and the theory of reduplication*. 28. Walter de Gruyter.

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

Andrew Spencer. 2005. Case in hindi. In *Proceedings of the LFG05 Conference*, pages 429–446. CSLI Publications Stanford, CA.

Jon Sprouse and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, 2:1–32.

Adrian Staub. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Andreas Stolcke. 2002. SRILM — An extensible language modeling toolkit. In *Proc. ICSLP-02*.

David Templin. 2013. The devanagari script. *Hindilangauge. info*.

Matthew Traxler and Morton Ann Gernsbacher. 2011. *Handbook of psycholinguistics*. Elsevier.

Appendix

Rating	Description
1	Bad or ungrammatical sentence
2	Poor Sentence
3	Fair Sentence
4	Good Sentence
5	Excellent Sentence

Table 4: Descriptions of quality ratings

Annotation Manual

This section describes our annotation manual, which serves as a comprehensive reference manual for annotators, while identifying errors, refining the corpus and assigning quality ratings to sentences. Each annotator was asked to encode the following 5 attributes as columns in an excel spreadsheet:

1. **Sentence ID**: Copy and paste the existing Sentence IDs from the Corpus.
2. **Quality rating**: Rate the sentences based on your understanding or speaker's intuition. Use a scale of 1 to 5 (see Table 4 for descriptions of each quality rating value)
3. **Remarks**: Provide details about what is wrong with the sentence.
4. **Raw sentence**: Paste the sentence as it is, without any modifications.
5. **Annotated sentence**: Correct the grammatical errors in the sentence while ensuring that the meaning and essence remain unchanged.

For the **Remarks** column, our annotators were asked to refer to the descriptions of various linguistic phenomena given below:

- CODE MIXING: The embedding of linguistic units such as phrases, words, and morphemes of one language into an utterance of another language.
- CODE SWITCHING: The term code-switching refers to a person changing languages or dialects throughout a single conversation and sometimes even over the course of a single sentence.
- MISSING/ADDITION OF AUXILIARY VERBS: Identify if auxiliary verbs are missing or need to be added.
- ADDITION OF WORD: If you believe there is an extra word that is not necessary.
- WORD REPETITION: Note instances when a sentence contains frequently repeated words.
- REDUPLICATION: Identify words formed through repetition of sounds or words.
- MISSING/ADDITION OF CASE MARKERS: Note any missing or incorrectly added cases.
- SENTENCE FRAGMENT: If the sentence is incomplete and constitutes only a part of a complete sentence.
- SENTENCE LENGTH: Indicate when a sentence is excessively long.
- AMBIGUOUS WORDS: Highlight words that have multiple meanings, leading to ambiguity.
- ABRUPT ENDING: Note if the sentence ends abruptly.
- WRONG SYMBOL: Identify spelling mistakes related to symbols in the transcribed speech text.
- ADDITION OF ELEMENTS: Note whether any additional pronouns, or adjectives, are required to complete the sentence.
- WRONG AGREEMENT: Highlight cases where there is incorrect agreement in terms of case or verb forms.