Forward Surprisal Models Production Planning in Reading Aloud

Sidharth Ranjan[1] Rajakrishnan Rajkumar[2] Sumeet Agarwal[1]
IIT Delhi[1], IISER Bhopal[2]
sidharth.ranjan03@gmail.com, rajak@iiserb.ac.in, sumeet@iitd.ac.in

The *Dual-Route-Cascaded* model (Coltheart et al., 2001, DRC) model of word recognition and reading aloud predicts that high-frequency words are read aloud faster than low-frequency words. In this work, we validate the aforementioned prediction using a publicly available read-aloud Hindi speech corpus (`https://tdil-dc.in`; Two speakers; S1: 341 sentences; 4,444 words and S2: 1,190 sentences; 11,163 words). We used the regression modelling framework proposed by Bell et al. (2009) and adapted their following bigram probability measure to capture *production planning* when reading aloud. We defined a measure named *forward trigram surprisal*, inspired from *Surprisal Theory* metric (Hale, 2001; Levy, 2008). Though originally proposed for comprehension, surprisal has been shown to correlate with word duration (Demberg et al., 2012) and disfluencies (Dammalapati et al., 2019) in spontaneous speech. We reason that if forward surprisal predicts word duration during reading aloud, we can conclude that some effect relevant to articulatory planning is captured by this measure. This is because readers might be incorporating parafoveal viewing in their planning.

We computed surprisal of the target word considering two kinds of contextual information in the sentence, viz. a) probability of the target word given the two previous words (*backward trigram surprisal*) and b) probability of target word given the two following words (*forward trigram surprisal*). We also calculate the probability of individual words (unigram surprisal) to capture frequency effects and test the DRC's prediction pertaining to frequent words in our dataset. In regard to the effect of letters on sound, Vaid and Gupta (2002) showed that Hindi graphemic complexity affects reading. Therefore, to investigate its effect on word duration, we calculated word length as the total number of vowels and consonants present in each word.

We trained a linear mixed-effects model containing word length, unigram, backward, and forward surprisal measures to predict word duration. We controlled for various random factors, including subject, item, parts of speech, and lexical class (content or function). As shown in Table 1, all four predictors have a positive coefficient and are significant predictors of word duration. This implies that an increase in predictor's score leads to higher reading time, indicating production difficulty. Moreover, the positive regression weight for unigram surprisal further validates the DRC prediction stated at the outset. Figure 1 shows the correlation coefficients among different predictors. Our results propound that despite the presence of unigram surprisal and word length predictors in the regression model, trigram surprisal measures are significant, denoting the effect of contextual predictability on word duration. Moreover, the significant effect of forward trigram surprisal suggests in particular that articulatory planning is incremental and continuous in nature, such that the planning of a target word is facilitated by both previous words and upcoming words, which are itself involved in subsequent planning (Pluymaekers et al., 2005). We also investigated the interaction between the predictors and found that the effect of forward trigram surprisal on reading times decreases by 0.02 with every unit increase in word length. This indicates that if the forward surprisal effect is driven by parafoveal preview, then the processor cannot compute the target word's surprisal given the longer following words as they would not fit the parafoveal window. Now we turn our discussion towards lexical classes viz., content and function words. Interestingly, Bell et al. (2009) showed that both backward and forward bigram predictability were significant predictors of word duration in spontaneous speech, but with asymmetric behavior towards content and function words such that they are accessed differently in production. However, we demonstrate that irrespective of the lexical class, all our predictors performed significantly. This suggests that during reading aloud processes, the access of lexical items to the extent of the full semantic representation of a word may not be necessary.

Predictors / Intercept	Estimate	Std. Error	t-value
Intercept	-1.864	0.071	-26.5***
Word length (Wordlen)	0.174	0.005	36.8***
Unigram surprisal (UniSurp)	0.049	0.006	8.2***
Backward Trigram surprisal (BckSurp)	0.015	0.003	5.7***
Forward Trigram surprisal (FwdSurp)	0.013	0.003	4.9***

Table 1: Regression model containing all predictors (15607 data points; all predictors significant with $p < 0.001$ denoted by ***)

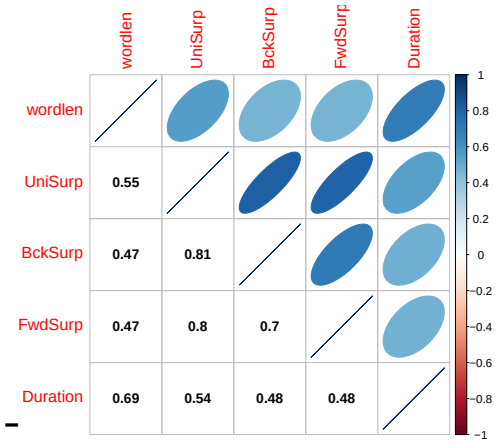

Figure 1: Correlation coefficients among different predictors and word duration

References

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., and Xia, F. (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.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Dammalapati, S., Rajkumar, R., and Agarwal, S. (2019). Expectation and locality effects in the prediction of disfluent fillers and repairs in english speech. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 103–109.

Demberg, V., Sayeed, A. B., Gorinski, P. J., and Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 356–367. Association for Computational Linguistics.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

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

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

Vaid, J. and Gupta, A. (2002). Exploring word recognition in a semi-alphabetic script: The case of devanagari. *Brain and Language*, 81:679–90.

Hindi

Hindi[1] is an Indo-Aryan language primarily spoken in India. It is a head final language with subject-object-verb (SOV) as canonical word-order and belongs to the Indo-European family. Hindi is considered a relatively free word-order language compared to English and offers a rich case-marking system using postpositions. The following sentence illustrates one such example taken from our dataset.

(1) telephone ek aadhunik *yug-ka* yantra hai
telephone one modern era-GENITIVE machine be.PRESENT.SINGULAR
The telephone is a modern era machine.

In the sentence example 1, we estimated the probability of each word given two previous words and two following words in context. For example, assuming our target word is "aadhunik", we estimate backward probability by taking the ratio of the frequency of sequence "*telephone ek aadhunik*" and frequency of "*telephone ek*". We estimate forward probability of target word "aadhunik" by taking the ratio of the frequency of sequence "*aadhunik yug ka*" and frequency of "*yug ka*". In order to compute corresponding surprisal scores, we took a negative logarithm value of these two contextual probabilities. However, for unigram probability, we calculate the frequency of each word present in our speech corpus data from EMILLE[2] Hindi corpus and then divided it by the total number of words in EMILLE corpus. The unigram surprisal was obtained by taking the negative logarithm value of the unigram probability.

The writing system of Hindi follows Devanagari[3] alphasyllabary based on Brahmi script. The Devanagari script is majorily composed of 47 characters containing 33 consonants (for e.g., क, ख, ग etc.) and 14 vowels (for e.g., अ, आ, इ etc.). Unlike Latin alphabet, Hindi has no concept of letter case (upper/lower) except for sinistrodextral (left-to-write) writing system. Each unit of word is written in horizontal direction separated by space and follows standard punctuation markers alike English except for full stop (.) where a pipe (।) is used as an end of sentence marker. In regards to letter-sound correspondence, the orthography of the script mostly corresponds with grapheme pronunciation except for cases when vowel diacritics, conjunct consonants or ligatures are present. Vowel diacritics (glyph) combines with consonants (क + अ = का) to form another syllabic letter. For example, the vowel – अ combines with consonant – क to give a letter का. Conjunct consonants is understood to offer most difficulty during reading consist of two consonants grouped together but with a missing vowel sound between them. For example, the two consonants (च, छ) when combined together (च + छ = च्छ), the letter च्छ (as in the word– अच्छा) has a missing vowel (अ) diacritic i.e., ◌् between them.

Therefore, to explore the effects of these graphemic complexities on reading read-aloud word duration, we included word length metric into our regression model. As shown in the literature, we also anticipate that increase in word length will increase the word duration. We calculated word length in two ways – firstly, counting the number of Unicode characters in Devanagari script, and secondly, the sum of the total number of consonants and vowels. For example, based on the aforementioned definitions of word length, word – इसलिए will have 5 Unicode characters (इ, स, ◌ि, ल, ए), 2 consonants (स, ल) and 2 vowels (इ, ए). Therefore, the two possible word lengths for the word – इसलिए are 5 and 4. In this study, however, we have reported results based on word length estimated using the total number of consonants and vowels present in the word as both the word length metrics are highly correlated (0.83). The correlation between word length and word duration comes out to be 0.69, as shown in Figure 1.

[1] https://en.wikipedia.org/wiki/Hindi
[2] https://www.lancaster.ac.uk/fass/projects/corpus/emille/
[3] https://en.wikipedia.org/wiki/Devanagari

Supplementary Materials

Surprisal Theory (Hale, 2001; Levy, 2008) posits that comprehenders build probabilistic knowledge based on previously experienced structures and upcoming linguistic information. It proposes a surprisal metric (an entity of mental surprise) to account for comprehension difficulty.

$$S_{k+1} = -\log P(w_{k+1}|w_{1...k}) = -\log \frac{P(w_1...w_{k+1})}{P(w_1...w_k)} = -\log P(w_{k+1}|T) = -\log \frac{\sum_T P(T, w_1...w_{k+1})}{\sum_T P(T, w_1...w_k)} \quad (1)$$

Mathematically, *surprisal* of the $(k+1)^{th}$ word, w, is defined as negative logarithm of conditional probability of word, w given the sentential context which can be either *sequence of words* or a *syntactic tree* (see Equation 1). All the surprisal measures investigated in this work were computed by training unigram (n=1) and trigram (n=3) language models on 1 million sentences of Hindi EMILLE corpus using the SRILM toolkit[4] with Good-Turing discounting. The per-word surprisal scores were used as an independent variable in the linear regression model to predict word duration. We observe that despite a very high correlation score (0.80) among unigram and trigram surprisal, the presence of both viz., backward and forward trigram surprisal in the model containing unigram surprisal (frequency effects) feature account for their combined as well as their individual effects validating the findings in the literature (Bell et al., 2009).

The word duration was extracted from the Hindi speech dataset using a software package PRAAT[5]. We trained random mixed-effect linear regression models on a dataset containing per-word duration, word length, and surprisal predictors. The dependent variable i.e, *duration* (Equation 2) was transformed into logarithmic scale. The logarithm scale of the independent variables viz., surprisal metrics, took care of extreme frequencies effects on the model computation. All the independent variables were scaled, i.e., the predictor's score (centered around its mean) was divided by its standard deviation. We have used Glm package in R to perform our regression experiment.

$$\textit{Duration} \sim \textit{word length} + \textit{unigram surprisal} + \textit{backward surprisal} + \textit{forward surprisal} \quad (2)$$

As discussed previously, we controlled for different parts of speech tags and lexical class viz., content, and function words, apart from subject and items in random mixed effects. Parts of speech tags for our dataset were obtained by training a Stanford parts-of-speech (POS) tagger (94% prediction accuracy) on publicly available human-annotated Hindi-Urdu Treebank (HUTB) corpus (Bhatt et al., 2009). After that, we automatically annotated each word in our dataset with lexical class viz., *content*, and *function* word annotation corresponding to the obtained POS tag of each word. Content word associates more with semantics, whereas function words with syntactic aspects in the sentence. The mean word length of a content word in the experimental items was 2.66 (minimum: 1, maximum: 8), and the function word was 1.74 (minimum: 1, maximum: 5).

Word length has been associated with duration by virtue of orthographic-phoneme correspondence (Vaid and Gupta, 2002). Thus, a word with longer syllable length will require more time to pronounce, leading to a longer word duration. The mean word length in the experimental items was 2.17 (minimum: 1, maximum: 8). Along these lines, we further investigated if *duration per character* can be predicted similarly, as shown for word duration (see Table 1) in our study. We transformed the response variable as duration divided by word length and regressed it over all the surprisal predictors except the word length metric. We found that the regression coefficient for forward and backward trigram surprisal remained positive, but it turned negative for unigram surprisal. This requires further experimentation using character-based language models that we shall pursue in future work.

[4] http://www.speech.sri.com/projects/srilm/

[5] http://www.fon.hum.uva.nl/praat/