Journal of Eye Movement Research
8(2):3, 1-12

Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus

Samar Husain
Indian Institute of Technology, Delhi, India

Shravan Vasishth
University of Potsdam, Germany

Narayanan Srinivasan
CBCS, University of Allahabad, India

This is the first attempt at characterizing reading difficulty in Hindi using naturally occurring sentences. We created the Potsdam-Allahabad Hindi Eyetracking Corpus by recording eye-movement data from 30 participants at the University of Allahabad, India. The target stimuli were 153 sentences selected from the beta version of the Hindi-Urdu treebank. We find that word- or low-level predictors (syllable length, unigram and bigram frequency) affect first-pass reading times, regression path duration, total reading time, and outgoing saccade length. An increase in syllable length results in longer fixations, and an increase in word unigram and bigram frequency leads to shorter fixations. Longer syllable length and higher frequency lead to longer outgoing saccades. We also find that two predictors of sentence comprehension difficulty, integration and storage cost, have an effect on reading difficulty. Integration cost (Gibson, 2000) was approximated by calculating the distance (in words) between a dependent and head; and storage cost (Gibson, 2000), which measures difficulty of maintaining predictions, was estimated by counting the number of predicted heads at each point in the sentence. We find that integration cost mainly affects outgoing saccade length, and storage cost affects total reading times and outgoing saccade length. Thus, word-level predictors have an effect in both early and late measures of reading time, while predictors of sentence comprehension difficulty tend to affect later measures. This is, to our knowledge, the first demonstration using eye-tracking that both integration and storage cost influence reading difficulty.

Keywords: reading, Hindi, eye-tracking, sentence comprehension, integration cost, storage cost

Introduction

Eyetracking corpora have been widely studied for languages such as English (Schilling, Rayner, & Chumbley, 1998; Kennedy, 2003) and German (Kliegl, Nuthmann, & Engbert, 2006). They have been used to study not only eye-movement control (Reichle, Rayner, & Pollatsek, 2004; Engbert, Nuthmann, Richter, & Kliegl, 2005), but also sentence processing difficulty, specifically, the predictions of computationally implemented theories such as surprisal, and working-memory based accounts (Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Boston, Hale, Vasishth, & Kliegl, 2011; Demberg & Keller, 2008). As such, these corpora are interesting for a wide range of disciplines, encompassing psychology, sentence comprehension research in psycholinguistics, and cognitive modeling.

Unfortunately, research on eyetracking corpora involving Asian languages is rare (exceptions are Chinese, e.g., Yan, Kliegl, Richter, Nuthmann, & Shu, 2010, and Uighur, Yan et al., 2014). In this paper, we present an analysis of an eyetracking corpus of Hindi that we have developed, the Potsdam-Allahabad Hindi Eyetracking Corpus. Our focus in this paper is on predictors of language processing difficulty as indexed by fixation-based measures.

Hindi is a language spoken primarily in India. It is difficult to estimate the number of speakers worldwide; one estimate is 180-258 million speakers (http://en.wikipedia.org/wiki/Hindi). Hindi belongs to the Indo-European family and is head-final; i.e., the default word order is subject-object-verb. It is characterized by relatively free word order and overt case-marking using postpositions.

The Hindi sentences used in the study have several attractive properties: the sentences used in the corpus are taken from the beta version of the Hindi-Urdu treebank (Bhatt et al., 2009), and are therefore already annotated for syntactic structure and part-of-speech. This allows us to compute several low-level (lexical-level) and high-level (sentence-level) predictors of reading difficulty. The corpus can therefore serve as a basis for investigating theories of eye-movement control and theories of sentence comprehension. It is intended to add to the existing large-scale naturalistic data-sets that are available for investigating theories of reading difficulty.

Our study represents a first attempt at characterizing reading difficulty in Hindi in naturally occurring sentences. We begin by explaining how the Hindi script (Devanagari) is structured; understanding the details of the script is important for the various word-level predictors we discuss. Then, we describe the various predictors of reading difficulty that were computed from the Hindi Treebank. We then provide statistical analyses using various reading time measures and outgoing saccade length as a dependent variable. In particular, the effect of the following predictors on reading difficulty is investigated: graphemic complexity, syllable length, unigram and bigram frequency, integration and storage cost.

Devanagari: The Hindi script

Hindi is written in the Devanagari script which has 13 vowels and 33 consonants, in addition, there are three consonant clusters with special symbols (Kachru, 2006). Vowels take different forms when they occur independently (eg. आ /a/) and when they appear with a consonant (eg. m + a = म + ा → मा /ma/). Conjunct consonants cạn sometime appear with a reduced form (eg. /s/ + /[thə/ = स् + थ → स्थ) or sometimes can take a different form (eg. /t/ + /rə/ = त् + र → त्र). A consonant character in Devanagari appears by default with a schwa sound /ə/ (eg. र /rə/). For more details see Kachru (2006),Wikipedia (2014).

Devanagari is read from left to right; words and case-marking morphemes are separated by spaces, and there is no upper- and lower-case distinction. Each word-unit that is separated by spaces usually has a horizontal line spanning the characters. Sentence-final full-stops are written as a vertical line, but standard punctuation markers such as commas are also used.

An interesting feature of Hindi orthography is that the linear position of a grapheme in the text does not always correspond to the order in which the graphemes are pronounced. For example, vowels can be written as diacritic symbols below or above a consonant but pronounced after the consonant; and short vowels can occur before a consonant even though they are pronounced after the consonant. In all such instances, when this asymmetry between writing order and pronunciation order occurs, it is possible that the difficulty in reading increases. Vaid and Gupta (2002) have investigated this issue, and found some evidence that mismatches between orthography and pronunciation impact reading of isolated words.

Method and Materials

Participants

Thirty graduate and undergraduate students of the University of Allahabad participated in the experiment for payment. All of them had had an Urdu medium education until at least high school and described themselves as fluent in reading both the Perso-Arabic script used for Urdu as well as the Devanagari script used for Hindi.

As noted above, the experiment was conducted in an urban (university) setting. Mono-lingual readers in India are rare (especially in an urban areas). Most educated individuals have considerable familiarity with more than one script. For example, English is taught in almost all schools (except remote areas where illiteracy is also an issue). Therefore familiarity with Latin script is common. Similarly, all college-going individuals have good command over the Latin script, as the medium of higher education in India is often English. The speakers who participated in this study could read Hindi (in Devanagari script) and Urdu (in Perso-Arabic script). In the part of India where this experiment was conducted, exposure to Devanagari script happens quite early in schooling as Hindi is a compulsory subject from pre-school until at least pre-college education. This holds true irrespective of whether the medium of instruction is Hindi, Urdu or English. In addition, individuals often need to know Hindi in order to negotiate their day-to-day activities; this is because road signs, advertisements, shop signs, etc., are often in Hindi. To summarize, in a setting where this experiment was conducted, mono-lingual readers are rare and finding such individuals is very difficult. We decided to make a virtue out of this difficulty by systematizing the collection of data in the two languages that readers were likely to be familiar with in that particular part of India (Allahabad). We do not report the Urdu data here as it would make the paper too complex. We plan to discuss the Urdu data in a separate paper.

Equipment

The experiment was conducted using the SMI iView X HED eyetracker with 500Hz sample rate. The subject was seated 50cm from the stimulus screen. Sentence were shown at the centre of the screen in a single line. The monitor used to display was Acer 19" LED with a 1600×900 screen resolution. The refresh rate of the monitor was 60Hz. Hindi text was displayed using the Mangal true-type 17 point font[1]. On average approximately 1.8 syllables[2] subtend 1° of visual angle in this experimental setup.

Materials

A subset of the Hindi-Urdu treebank data (Bhatt et al., 2009) which has 400,000 words was used to get the

[1] This proportional font type used in the experiment is supplied by Microsoft and therefore is quite extensively used.

[2] A consonant-vowel combination is considered a syllable. On average 1.6 characters form a syllable in the data. See section on 'syllable length' for more details.

experimental sentences. We transcribed the Hindi data using the Perso-Arabic script to get the Urdu sentences. This gave us identical text in two different scripts for the two languages. This provided us an opportunity to study reading processes in both languages using our bilingual subject pool. As noted earlier, while Hindi is written in the Devanagari script, Urdu is written in the Perso-Arabic script. Structurally Hindi and Urdu are almost identical; however some differences exist in the lexicon. Hindi/Urdu spoken colloquially have a shared vocabulary as well. Since Hindi treebank text had some Sanskritized words, which participants may not be familiar with, these were substituted with more colloquial alternatives. We used 153 sentences (2610 words) for each language (Hindi and Urdu), and additionally four sentences were used as practice sentences. We avoided using sentences that had a political bent. The target sentences that were chosen were about topics such as movies, entertainment, and sports. The target sentences chosen were not isolated sentences, but formed short narratives consisting of several sentences. Each sentence from a narrative was presented separately on a screen, and the end of a narrative was signaled by a blank screen.

Procedure

Participants were required to read identical texts in Hindi (Devanagari) script and Urdu (Perso-Arabic) script. Since the content of the two scripts was identical, the experiment was conducted in two blocks over two days. The order of presentation was pseudo-randomized such that participants were exposed to one of eight combinations of these two factors. Table 1 shows all the groups. Each participant was randomly assigned to one of these groups. For example, in Group 1, the first part of Hindi text (74 sentences) was read in the first block of the first session, then after an interval of 5 minutes, the second part of the Urdu text (79 sentences) was read in the second block. In order to reduce the effect of familiarity, participants read the remaining sentences of each language after a few days. The reading task on the second day also consisted of two blocks. The average gap between the two sessions was 5.7 days. One concern with such a setup could be that the text read in the second session will be influenced by the text read in the first session. We therefore report the results by using session id as a factor in the final analysis to determine whether this had an effect; all interactions with session were also investigated.

The experiment started with the experimenter orally briefing the participant as regards the task. This was followed by subject reading written instructions on the computer screen. Following this, a 13-point calibration was performed. The experiment started with four practice sentences, following which the experimental sentences were presented. A trial started with the presentation of a gaze-correction point on the centre left of the screen. Fixating on this point briefly led to the presentation of the sentences. After reading the sentence, the participant looked at a small dot on the bottom-right of the screen and pressed the left-button of a mouse. Recalibration was done after every 15 sentences or if the fixation on the gaze-correction point didn't trigger the sentence presentation. A blank screen was presented to signal the end of a narrative.

Table 1
Experiment session groups.

Group	Session 1		Session 2	
	Block 1	Block 2	Block 1	Block 2
Group 1	H1	U2	U1	H2
Group 2	U1	H2	H1	U2
Group 3	H2	U1	U2	H1
Group 4	U2	H1	H2	U1
Group 5	H1	U2	H2	U1
Group 6	U1	H2	U2	H1
Group 7	H2	U1	H1	U2
Group 8	U2	H1	U1	H2

Notes: *H and U stand for Hindi and Urdu respectively. H1 and U1 comprised of 74 sentences, while H2 and U2 had 79 sentences.*

Comprehension questions were not asked due to time constraints. Although participants were instructed to read the sentences carefully so that they understand its meaning, it is quite possible that they did not do so while reading a sentence. However, the results show that the participants were indeed attending to the sentences carefully. Further evidence comes from the results that are consistent with reading patterns in other languages. For example, the effect of word (syllable) length and word frequency is consistent with previous literature. In addition to this, we see a significant effect of sentence-level processing factors such as storage cost and distance cost; this suggest active involvement of the subjects during the reading process.

Computing word and sentence level predictors

We computed several measures of processing difficulty for this corpus. It is well-known in the eye-movement research literature that word length, and unigram and bigram frequency are predictors of reading difficulty (Rayner, 1998; McDonald & Shillcock, 2003; Kliegl et al., 2006). In addition, due to the special properties of Devanagari characters, we also developed a metric for graphemic complexity. We also computed a metric for sentence comprehension difficulty based on the work of Gibson (2000); distance cost and storage cost. Together, these predictors can be seen as representative of so-called low-level and high-level predictors of processing difficulty (Boston et al., 2008, 2011; Demberg & Keller, 2008). Table 2 shows a summary of the distributions of the predictors.

A metric for word complexity. In the appendix we present a first attempt at quantifying the complexity cost of Hindi characters. The work by Vaid and Gupta (2002) on the effect of character complexity on reading served as a guide when developing this metric. In essence, our metric defines a linear penalty metric for mismatches in character order and pronunciation order: (a) if a vowel diacritic appears to the left of the consonant but is pronounced after the consonant, the cost is 1; (b) if a diacritic appears above or below a consonant, the cost is 0.5, (c) if a consonant appears in a consonant cluster, i.e. without its inherent vowel, the cost is 0.5, and (d) ligatures get a cost of 1. The assumption here is that violation of character order (relative to pronunciation order) should get the maximum penalty because that seems to be the cause of greatest complexity (Vaid & Gupta, 2002); diacritics and consonants without vowels do not violate order as such, but they do require more processing effort than the cases where character order matches pronunciation order perfectly. Under this metric, the mean word complexity in the Hindi text was 0.46 (minimum: 0, maximum: 5.5). We also experimented with a metric that penalizes all deviations from the simplest case equally; the results were comparable to the one reported using the metric described above.

Syllable length. The syllable boundary is used for computing word length, in particular, a consonant-vowel combination is considered a single unit. For example, मिल /mɪl/ has a syllable length of 2 = 1 (मि) + 1 (ल). In case of ligatures leading to complex forms or for composite character, the entire combination is considered as a single unit, for example, the syllable length of कार्निवाल /kɑrnɪval/ will be 4 = 1 (का) + 1 (र्नि) + 1 (वा) + 1 (ल). Likewise, the syllable length of प्रधानमंत्री /prəd^{h}anməntri/ will be 5 = 1 (प्र) + 1 (धा) + 1 (न) + 1 (मं) + 1 (त्री). This criterion for segmentation is also influenced by practical concerns of the eyetracking paradigm. It would be difficult to ascertain the gaze position accurately at the level of the individual character especially in cases such as discussed above. The mean syllable length in the experimental items was 2.2 (minimum: 1, maximum: 10).

Unlike character-based scripts such as Latin, in Devanagari, a consonant or a vowel need not take constant space. In addition, as stated above the characters combine to form ligatures; they also appear as diacritics above or below another character. Given these properties, we found it reasonable to compute word length based on syllable count. We also computed the standard definition of computing word length, i.e., counting the number of consonants and vowels in a word. Word length computed using this standard definition is correlated (.60) with graphemic complexity. The results obtained using this definition were similar to the one obtained using the syllable-based word length.

Frequency (unigram and bigram). The unigram and the bigram frequencies were computed using the beta version of the Hindi-Urdu treebank data (Bhatt et al., 2009), which has 400,000 words.

The mean token frequency was 3837 (minimum: 1, maximum: 19420); while mean type frequency was 6915 (minimum: 1, maximum: 25350). The mean bigram frequency (token) was 177.8 (minimum: 1, maximum: 6561); while mean bigram frequency (type) was 302.3 (minimum: 1, maximum: 8730). Since token and type frequencies are highly correlated (cf. Table 3), we only use token frequencies as predictors in the analysis.

Distance cost. Integration cost is a processing metric proposed by Gibson (2000) as part of a more general Dependency Locality Theory (DLT). It intends to capture the retrieval cost of a dependent at its integration site (also see Lewis & Vasishth, 2005); in other words, the integration cost metric aims to characterize the online processing cost of completing the dependency link between an already seen/heard word and co-dependent being currently processed. Some examples are subject-verb dependencies, and antecedent-reflexive dependencies. We computed an approximation of integration cost: the distance in words between two co-dependents. For example (1), the distance cost at 'narrated' would be 8 = 5 (for दीपिका) (ne, Abhay, ko, ek, kahaanii), and 3 (for अभय) (ko, ek, kahaanii). The distance cost was calculated manually; we did not compute dependencies using a dependency grammar representation because we wanted to ensure that there was no loss of accuracy.

(1) a. दीपिका ने अभय को एक कहानी
Deepika ERG Abhay DAT a story
सुनाई
narrated
'Deepika narrated a story to Abhay.'

The mean distance cost was 1.15 (minimum: 0, maximum: 66, mean 1.3, sd 4.7). The distribution of distance scores is well-modeled by an exponential distribution with rate 0.77.

Storage cost. While integration cost is intended to characterize the cost of completing a dependency, storage cost was proposed by Gibson (2000) to characterize the processing load incurred as a result of maintaining predictions of upcoming heads. In example (1), the storage cost at the verbal arguments (दीपिका, अभय and कहानी) would be 1, while the storage cost at the verb is 0. The mean storage cost was 1.01 (minimum: 0, maximum: 3). Storage cost was also computed by hand.

The correlations between the predictors are shown in Table 3. As expected, syllable length and frequency are negatively correlated (−0.63), word frequency and bigram frequency have correlation 0.36. Distance cost

Table 2
Minimum, first quartile, median, mean, third quartile and maximum values of all the predictors.

	Minimum	First quartile	Median	Mean	Third quartile	Maximum
syll_len	1	1	2	2.2	3	10
word_complex	0	0	0.5	0.4	1	5.5
word_freq	1	27	395	3837	5500	19420
word_bifreq	1	1	3	179.3	25	6561
IC	0	0	0	1.1	0	66
SC	0	1	1	1	1	3

Notes: *The abbreviations have the following meaning: syll_len: syllable length; word_complex: word complexity; word_freq: word unigram frequency; word_bifreq: word bigram frequency; IC: integration cost; SC: storage cost.*

and storage cost are negatively correlated -0.30; this is plausible: the longer dependent-head distances would occur in cases where most of the heads have already been seen.

Results

Statistical analyses

All analyses for fixation measures were carried out with Bayesian linear mixed models using Stan, version 2.5 (Stan Development Team, 2014). We fit full variance-covariance matrices for the subject- and item-level main effects and interactions, including correlation estimates (i.e., we fit two 14×14 variance-covariance matrices for subject and item effects, respectively). One of the advantages of using Bayesian hierarchical models rather than frequentist ones is that we can directly compute the posterior probability of the coefficient of a particular effect being positive or negative given the data; unlike the frequentist approach, there is no need to indirectly draw inferences about the effect by appealing to the questionable procedure of rejecting a null hypothesis and computing a p-value (see, for example, Gelman (2013)). Another advantage is that we can fit a statistical model that takes into account all possibly relevant variance components. This currently cannot be done with the frequentist tools available, because of convergence or estimation failures. Bayesian hierarchical models do not suffer from this problem because mildly informative priors are defined over all parameter estimates; if there is insufficient data to estimate the parameters, the prior will dominate in determining the posterior distribution, and will ensure that the posterior mean is near 0.

The details of the Bayesian model-fitting procedure are discussed in detail in (Sorensen & Vasishth, 2014) and in the R package RePsychLing, available on github. The source code for the models fit in the present paper is available from https://github.com/vasishth/StanJAGSexamples. The Stan analyses are summarized in the tables below using means and 95% posterior credible intervals for each coefficient. Credible intervals present the bounds within which we can be 95% certain that the true value of the parameter lies (given our particular data). We assume that an effect is present if the 0 value is not within the 95% credible interval.

All predictors were scaled; each predictor vector (centered around its mean) was divided by its standard deviation. Saccade and fixation detection was done using the saccades package developed by von der Malsburg (https://github.com/tmalsburg/saccades). Fixation measures were computed using the R package `em2` (Logačev & Vasishth, 2014) (downloadable from http://cran.r-project.org/src/contrib/Archive/em2/). We present analyses for one representative first-pass measure, first-pass reading time, and two representative measures that often show the effects of sentence comprehension difficulty, regression-path duration and total reading time (Clifton, Staub, & Rayner, 2007; Vasishth, von der Malsburg, & Engelmann, 2012). First-pass reading time on a word refers to the sum of the fixation durations on the word after it has been fixated after an incoming saccade from the left, until the word is exited to the right. Regression path duration on a word refers to the sum of the first-pass reading times and all fixations on preceding words, until the word is exited to the right. Total reading time is the sum of all fixations on a word; in other words, it is the sum of first-pass reading times and re-reading times. Each word served as a region of interest. All data points recorded with zero ms for these fixation measure (about 25% of the data) were removed, and the data analysis was done on log-transformed reading times to achieve approximate normality of residuals. Most of the zero ms fixations were due to short words being skipped entirely; this is quite normal in eyetracking data.

We also computed the length of the outgoing saccade (in syllables) from each word. This is defined as the length of a rightward saccade from a given word to a subsequent word during any pass, first-pass, or a revisit. The distribution of the saccade lengths can be modeled as an exponential distribution, with rate 0.36. Minimum

Table 3
The upper triangular correlation matrix for the predictors.

	word_complex	word_freq	type_freq	word_bifreq	type_bifreq	word_len	IC	SC
syll_len	0.35	-0.63	-0.69	-0.20	-0.23	0.85	0.02	0.08
word_complex		-0.19	-0.14	-0.10	-0.14	0.65	-0.03	0.02
word_freq			0.84	0.36	0.45	-0.58	-0.12	-0.16
type_freq				0.21	0.25	-0.56	-0.08	0.01
word_bifreq					0.88	-0.24	-0.06	-0.23
type_bifreq						-0.27	-0.06	-0.29
word_len							0.03	0.08
IC								-0.30

Notes: *The abbreviations have the following meaning: syll_len: syllable length; word_complex: word complexity; word_freq: word unigram frequency; type_freq: type unigram frequency; word_bifreq: word bigram frequency; type_bifreq: type bigram frequency; word_len: word length; IC: integration cost; SC: storage cost.*

outgoing saccade length was 1 and maximum 60, with mean 3 and sd 1.7. We used log saccade length as a dependent variable to investigate whether our predictors could influence saccade length. Although saccade length is not standardly used in sentence comprehension research, there is evidence that reduced processing difficulty could lead to longer outgoing saccade lengths (eg. Jacobson and Dodwell (1979); Rayner and Pollatsek (1989); Rayner, Ashby, Pollatsek, and Reichle (2004); White and Liversedge (2006); Wei, Li, and Pollatsek (2013)). Moreover, it is well known at least since Rayner (1979) that the length of an outgoing saccade depends partly on the length of the word fixated next; this is because the reader attempts to direct the saccade to the preferred viewing location of the next word. This preferred viewing position is slightly to the left of the center of a word. There is of course much more to be said about constraints on saccade launch and landing; but since our primary interest is in measures of sentence comprehension difficulty, we do not discuss these details any further.

Reading time and outgoing saccade length analysis

In log first pass reading times, we see effects of syllable length and bigram frequency in the expected directions: increase in syllable length leads to slower reading times, and higher bigram frequency leads to faster reading times. The credible intervals for unigram frequency include 0, but the posterior probability of the coefficient for frequency being less than 0 is 0.79. The distance cost metrics of integration cost and storage also have credible intervals including 0; the posterior probability of the IC coefficient being positive is 0.88, and of the SC coefficient is 0.67. Thus, there is only weak evidence for distance cost playing a role even in this relatively early measure of reading difficulty. Finally, although the credible interval for the effect of session includes 0, the posterior probability of the coefficient for session being less than 0 is 0.94; in other words, in the second session, readers tended to read faster. None of the interactions between session and the other factors seem have a large effect.

In log regression path durations, we see effects of syllable length and bigram frequency in the expected directions. The credible intervals for all other predictors include 0. Perhaps surprisingly, the coefficient for storage cost is negative, with a posterior probability of the coefficient being negative being 0.91. Thus, in log regression-path duration, we see *faster* reading times with increasing storage cost. We return to this point in the general discussion.

In log total reading time, we see effects of syllable length, unigram and bigram frequency, in the expected directions. In addition, we see an effect of storage cost, with higher cost leading to longer log total reading time. There is evidence for a session effect as well, with the second session leading to faster log reading time. None of the interactions between session and the other predictors seem to be relevant.

In log outgoing saccade length, we find effects of syllable length and unigram and bigram frequency. The effect of syllable length of the current word on log saccade length is consistent with the findings reported by Rayner (1979). As expected, the length (in syllables) of the word fixated next also has an effect: the outgoing saccade length is longer if the word fixated next is longer. This is due to the preferred viewing location effect discussed earlier. Regarding the syntactic distance measures, increasing integration cost leads to shorter saccade length, and increasing storage cost leads to longer saccade length. No effect of session seems to be present, and no interactions between session and the other predictors appears to have an impact.

Table 4
The effect of the predictors on log first-pass reading time, regression path duration.

	Log first-pass reading time		
	mean	lower	upper
Int	5.5019	5.4525	5.5507
sl	**0.1142**	**0.0878**	**0.1417**
comp	0.0002	-0.0100	0.0100
freq	-0.0055	-0.0189	0.0078
bifreq	**-0.0124**	**-0.0206**	**-0.0041**
IC	0.0068	-0.0047	0.0181
SC	0.0029	-0.0102	0.0156
session	-0.0156	-0.0360	0.0049
sl x session	-0.0046	-0.0125	0.0031
comp x session	0.0003	-0.0042	0.0049
freq x session	0.0052	-0.0037	0.0145
bigram x session	-0.0005	-0.0055	0.0045
IC x session	-0.0018	-0.0098	0.0060
SC x session	-0.0000	-0.0061	0.0060
	Log regression path duration		
	mean	lower	upper
Int	5.6540	5.5905	5.7175
sl	**0.1238**	**0.0961**	**0.1504**
comp	-0.0005	-0.0143	0.0129
freq	-0.0064	-0.0245	0.0117
bifreq	**-0.0225**	**-0.0321**	**-0.0136**
IC	0.0129	-0.0021	0.0275
SC	-0.0117	-0.0289	0.0052
session	**-0.0299**	**-0.0499**	**-0.0101**
sl x session	-0.0055	-0.0153	0.0044
comp x session	0.0033	-0.0029	0.0096
freq x session	0.0032	-0.0067	0.0130
bigram x session	-0.0020	-0.0080	0.0041
IC x session	-0.0082	-0.0189	0.0032
SC x session	0.0006	-0.0077	0.0092

Notes: *The columns present the results of the Bayesian hierarchical linear models; we show the estimated mean effect of each predictor, along with 95% credible intervals. All effects that have intervals excluding 0 are in bold. Int: intercept; sl: syllable length; comp: word complexity; freq: word unigram frequency; bifreq: word bigram frequency; IC: integration cost; SC: storage cost; session: session id.*

Discussion

To summarize the results, in log first pass reading times we primarily see stronger effects of “low-level” predictors than for syntactic-level processing difficulty such as integration cost; we also see some weak evidence for a session effect, with the second session showing faster reading times. In log regression path duration, we see clear effects of syllable length and frequency, and weak evidence for faster reading time with increasing storage cost. Log total reading time shows effects of syllable length and frequency in the expected directions, with an effect of storage cost, such that increasing SC results in longer reading times. Session effects are also seen: the second session is read faster. Finally, consistent with previous work on reading, log outgoing saccade length shows effects of syllable length: longer syllable length leads to longer log saccade length. Frequency also shows a clear effect: increasing frequency leads to longer outgoing saccades. Finally, increased integration cost leads to shorter saccade length.

The effects of low-level predictors on reading times are consistent with the findings in the literature on reading: longer syllable length leads to longer fixations, and higher frequency leads to shorter fixations. Perhaps surprisingly, we don’t find a reliable effect of graphemic complexity on reading difficulty. This is surprising because Vaid and Gupta (2002) did find effects of graphemic complexity. However, this absence of an effect may be due to several reasons. First, Vaid and

Table 5
The effect of the predictors on total reading time, and outgoing saccade length.

	Log total reading time		
	mean	lower	upper
Int	5.6138	5.5462	5.6799
sl	**0.1378**	**0.1104**	**0.1657**
comp	0.0014	-0.0116	0.0144
freq	**-0.0193**	**-0.0366**	**-0.0016**
bifreq	**-0.0206**	**-0.0339**	**-0.0092**
IC	-0.0023	-0.0154	0.0102
SC	**0.0185**	**0.0035**	**0.0333**
session	**-0.0287**	**-0.0504**	**-0.0068**
sl x session	-0.0076	-0.0161	0.0008
comp x session	0.0032	-0.0021	0.0084
freq x session	0.0072	-0.0036	0.0183
bigram x session	-0.0013	-0.0070	0.0044
IC x session	-0.0008	-0.0074	0.0060
SC x session	-0.0006	-0.0088	0.0077
	Log outgoing saccade length		
	mean	lower	upper
Int	0.9254	0.8352	1.0148
sl	**0.0732**	**0.0592**	**0.0877**
targetsl	**0.0795**	**0.0670**	**0.0922**
comp	-0.0010	-0.0108	0.0084
freq	**0.0349**	**0.0241**	**0.0449**
bifreq	**0.0113**	**0.0027**	**0.0198**
IC	**-0.0355**	**-0.0479**	**-0.0231**
SC	**0.0268**	**0.0156**	**0.0380**
session	0.0208	-0.0005	0.0430
sl x session	0.0022	-0.0037	0.0082
targetsl x session	-0.0013	-0.0064	0.0037
comp x session	-0.0013	-0.0067	0.0040
freq x session	0.0003	-0.0061	0.0064
bigram x session	-0.0002	-0.0053	0.0050
IC x session	-0.0017	-0.0065	0.0036
SC x session	0.0008	-0.0041	0.0059

Notes: *The columns present the results of the Bayesian hierarchical linear models; we show the estimated mean effect of each predictor, along with 95% credible intervals. All effects that have intervals excluding 0 are in bold. Int: intercept; sl: syllable length; targetsl: syllable length of the word fixated after the outgoing saccade; comp: word complexity; freq: word unigram frequency; bifreq: word bigram frequency; IC: integration cost; SC: storage cost; session: session id.*

Gupta did not test natural reading, but rather presented isolated words to subjects to read out. It is possible that in natural reading, readers process complex graphemes as a unit and are not affected by mismatches between character order and pronunciation order. A second possibility is that our graphemic complexity metric may not characterize the sources of difficulty correctly. A third possibility is that it may simply be a question of low statistical power. A larger scale study can clarify this point.

The effects of increasing word frequency on saccade length are as expected: increasing frequency (unigram and bigram) leads to longer outgoing saccades. This frequency effect is easily explained: higher frequency translates to greater processing ease, which may allow the current fixation to process more letters, thereby allowing a saccade to be programmed further to the right (Rayner et al., 2004), (Wei et al., 2013).

The effects of sentence-level processing difficulty are discussed next. We see reliable effects of dependency-head distance (integration cost) in log outgoing saccade length, but only weak evidence for this complexity metric in the reading time measures. The effect of integration distance on outgoing saccade length is perhaps not surprising: increased distance cost represents greater integration difficulty, which could lead to shorter outward saccades due to greater processing load. Storage cost shows an effect in log total reading times and outgoing saccade length; increased storage cost leads to longer total reading times, and longer outgoing saccades. Since no effect was seen in first-pass reading time, the total reading time result suggests that the storage cost effect is driven by re-reading times. In other words, it seems to be a late-emerging effect. It is difficult to be certain that storage cost does not have any effect in early measures such as first-pass reading time; it is possible that we failed to find a storage effect in these measures due to the relatively small sample size (30 participants; compare this to the Potsdam Sentence Corpus of Kliegl et al. (2006), which had over 200 participants). With a larger sample size, storage cost may well have an effect on early measures. It is interesting that increased storage cost leads to *longer* outgoing saccades. Although speculative, one possible explanation for this result could be that increased storage cost encourages the reader to look further to the right in order to verify whether the predicted head appears further downstream. This is a possibility worth investigating in a planned experiment.

General Discussion

This study reveals several interesting facts about Hindi sentence comprehension difficulty. A new result, not noticed in previous work on eyetracking corpora from other languages, is that both integration and storage cost impact reading difficulty, but only when we consider so-called late measures (regression-path duration and total reading time) and outgoing saccade length; we did not find strong evidence that the early measure, first-pass reading time, is affected by these variables.

Integration cost estimates the difficulty with which co-dependents are integrated while parsing a sentence. A standard assumption, going back to Just and Carpenter (1992) but more fully worked out by Gibson (2000), and Lewis and Vasishth (2005), is that the greater the dependent-head distance, the greater the difficulty in completing the dependency. The cause for this so-called locality effect could lie in decay (this is how the Dependency Locality Theory explain this, see Gibson, 2000), or in interference or some combination of interference and decay (this is how the cue-based retrieval model of Lewis & Vasishth, 2005 explains it; also see Lewis, 1996). Whatever the underlying explanation, there is clear evidence for locality effects in planned experiments (e.g., Grodner & Gibson, 2005; Bartek, Lewis, Vasishth, & Smith, 2011). However, there are several important counterexamples too; examples are the German studies done by Konieczny (2000), and the experiments involving Hindi by Vasishth and Lewis (2006). Konieczny suggests a variant of the idea that delaying the appearance of a head (effectively increasing head-dependent distance) can facilitate processing if the conditional probability of the head appearing increases with distance (Levy, 2008). The Vasishth and Lewis proposal is that if the intervening material activates the upcoming head, the dependent-head integration could be facilitated to the head being reactivated. It has been suggested by Levy, Fedorenko, and Gibson (2013) that these so-called anti-locality effects may be restricted to head-final languages. Our results show that, while that could be correct, at least in the present Hindi data, when dependency distance is increased, there is some evidence that processing difficulty generally increases.

The effect of storage cost is also quite interesting. Storage cost characterizes the effort required to maintain predictions of upcoming heads. For example, when reading a main clause, readers may predict an upcoming verb (a storage cost of 1). If a sentence with an embedded clause is read, then the reader would predict two heads (one for the embedded clause, and the other for the main clause), leading to a storage cost of 2. Although some evidence does exist for storage cost (Chen, Gibson, & Wolf, 2005), the present work may be the first eyetracking study using naturally-occurring sentences that investigates this metric. The evidence in favor of storage cost has interesting implications for theories of expectation-based processing. The current view in the field of sentence processing is that the dominant predictor of expectation cost is surprisal: the conditional probability of an upcoming part of speech or word given the left context (Hale, 2001). Our study shows that, at least in this head-final language, the number of expected heads may also play a role. An obvious question this raises is whether surprisal-based expectation has a larger effect size than integration- and storage-cost effects. To answer this question, a probabilistic parser needs to be developed for Hindi, and the surprisal metric computed. This would allow us to investigate the relative effect size of storage vs surprisal cost. We expect to take up this and other issues in future work.

Conclusions

This is, to our knowledge, the first study of Hindi sentence processing difficulty using an eyetracking corpus containing naturally occurring text. We show that the standard so-called "low-level" predictors influence reading time in the expected manner. In addition, we show that two "high-level" predictors of sentence comprehension difficulty, integration and storage cost, also affect reading difficulty. The timing with which low-level and high-level predictors impact reading difficulty seem to differ in Hindi: first-pass reading difficulty shows effects only of low-level predictors, while regression-path duration and total reading time show effects due to both low- and high-level predictors. Outgoing saccade length is also affected by low- and high-level predictors.

Acknowledgements

This research was funded by the University of Potsdam, Germany. We benefitted from comments by Lena Jäger, Sol Lago, Felix Engelmann, and the reviewers; in particular, we would like to thank Dr. Ehab Hermena for his careful reading of the manuscript and for his comments and suggestions for improvement. We thank the developers of the Hindi-Urdu treebank for allowing us to use the beta version of the Hindi-Urdu treebank.

References

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*(5), 1178–1198.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D., & Xia, F. (2009). A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu. In *Proceedings of the Third LAW* (pp. 186–189).

Boston, M. F., Hale, J. T., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12.

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*(3), 301–349.

Chen, E., Gibson, E., & Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, *52*(1), 144–169.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye Movements in Reading Words and Sentences. In R. V. Gompel, M. Fisher, W. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (chap. 15). Elsevier.

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

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*, 777-813.

Gelman, A. (2013). P values and statistical practice. *Epidemiology*, *44*, 69–72.

Gibson, E. (2000). Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, Language, brain: Papers from the First Mind Articulation Project Symposium.* Cambridge, MA: MIT Press.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, *29*, 261–290.

Hale, J. T. (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.* Pittsburgh, PA.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PLoS ONE*, *9*(7), 1–14.

Jacobson, J., & Dodwell, P. (1979). Saccadic eye movements during reading. *Brain and Language*, *8*, 303–314.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 122–149.

Kachru, Y. (2006). *Hindi.* John Benjamins Publishing Company, Philadelphia.

Kennedy, A. (2003). The Dundee Corpus [CD-ROM] [Computer software manual]. The University of Dundee, Psychology Department, Dundee, UK.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of experimental psychology: General*, *135*(1), 12.

Konieczny, L. (2000). Locality and parsing complexity. , *29(6)*, 627–645.

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

Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *Journal of memory and language*, *69*(4), 461–495.

Levy, R., & Keller, F. (2012). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*.

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25(1)*, 93–115.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45.

Logačev, P., & Vasishth, S. (2014). The em2 package for computing eyetracking measures [Computer software manual]. Potsdam, Germany.

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*, 1735-1751.

Nicenboim, B., Vasishth, S., Kliegl, R., Gattei, C., & Sigman, M. (2014). *Individual differences in long distance dependency resolution.* (submitted)

Radach, R., & McConkie, G. W. (1998). Determinants of fixation positions in words during reading. *Eye guidance in reading and scene perception*, 77–100.

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception*, *8*(1), 21–30.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of word frequency and predictability on eye movements in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 720–732.

Rayner, K., & Pollatsek, A. (1989). *The Psychology of Reading.* Englewood Cliffs.

Reichle, E., Rayner, K., & Pollatsek, A. (2004). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(04), 445–476.

Schilling, H., Rayner, K., & Chumbley, J. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory and Cognition*, *26*(6), 1270–1281.

Sorensen, T., & Vasishth, S. (2014). *A tutorial on fitting Bayesian linear mixed models using Stan.* (Unpublished manuscript)

Stan Development Team. (2014). Stan modeling language users guide and reference manual, version 2.2 [Computer software manual]. Retrieved from `http://mc-stan.org/`

Vaid, J., & Gupta, A. (2002). Exploring Word Recognition in a Semi-Alphabetic Script: The Case of Devanagari. *Brain and Language*, *81*, 679-690.

Vasishth, S. (2003). *Working memory in sentence comprehension: Processing Hindi center embeddings.* New York: Garland Press. (Published in the Garland series Outstanding Dissertations in Linguistics, edited by Laurence Horn)

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767-794.

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2012). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, 125–134.

Wei, W., Li, X., & Pollatsek, A. (2013). Word properties of a fixated region affect outgoing saccade length in chinese reading. *Vision research*, *80*, 1–6.

White, S. J., & Liversedge, S. P. (2006). Foveal processing difficulty does not modulate non-foveal orthographic influences on fixation positions. *Vision Research*, *46*, 426–437.

Wikipedia. (2014, September). *Devanagari.* Retrieved from `http://en.wikipedia.org/wiki/Devanagari`

Yan, M., Kliegl, R., Richter, E. M., Nuthmann, A., & Shu, H. (2010). Flexible saccade-target selection in Chinese reading. *The Quarterly Journal of Experimental Psychology*, *63*(4), 705–725.

Yan, M., Zhou, W., Shu, H., Yusupu, R., Miao, D., Krügel, A., & Kliegl, R. (2014). Eye movements guided by morphological structure: Evidence from the Uighur language. *Cognition*, *132*(2), 181–215.

Appendix A
Details of the word complexity metric

1. **Vowel diacritic appears to the left of a consonant**: The short, unrounded, high front vowel (/i/ इ) when appearing with a consonant is represented as a diacritic ि◌ and pre̩ce̩des the consonant in the text, for example, in दिन /[dı[n/ the vowel ि◌ /I/ precedes the consonant द /[də/ but is pronounced after the consonant. In effect, the written vowel appears displaced with respect to the point of it utterance. In related work, Vaid and Gupta (2002) found that words with such vowels lead to slower naming latencies and higher naming errors compared to control. In all such cases we posit a compl̩ex̩ity cost of 1; for example, the complexity cost of दिन /[dı[n/ would be 1.

In addition there is also a cost for the distance of displacement of the vowel, for example in पब्लिसिटी /pəblIsIʈi/ there is an additional consonant ब् /b/ between the vowel ि◌ /I/ and the consonant its associated with ल /lə/. In such cases the cost becomes 1+d, where d is the number of intervening consonants between the vowel and the consonant its associated with. Note that this situation will happen in cases where the preceding consonant appears without its inherent vowel. So the total cost for a word like पब्लिसिटी would be 3.5 = 2 (for the ि◌ in ब्लि) + 1 (for the ि◌ in सि) + .5 (for the ब्; see below)

2. **Diacritic above a consonant**: Although all vowels in Hindi have an independent form, when they combine with a consonant some of them can appear above the consonant. In all such cases (see, table A1) we assume a complexity cost of .5.

Table A1
Diacritics appearing above a consonant

◌े /e/	◌ै /ɛ/	◌ँ /˜/	◌ं /m/	◌ॅ /o/	◌ॉ /ɐ/

3. **Diacritic below a consonant**: Similar to the vowel that can appear above a consonant, some vowels can appear below a consonant. In all such cases (see, table A2) we assume a complexity cost of .5. Note that some consonants are Perso-Arabic borrowings, and in those cases the diacritic ◌़ is added to already existing letters. These are क़ /q/, ख़ /x/, ग़ /ɣ/, ज़ /z/, फ़ /f/ (Kachru, 2006).

Table A2
Diacritics appearing below a consonant

◌ु /u/	◌ू /ʊ/	◌़ (see footnote 1)

4. **Consonant without inherent vowel**: Consonants, when occurring without the inherent vowel, are written with a slightly different form, in many cases the vertical bar associated with the consonant is missing (eg. घ् + ट → घ्ट /g^hʈ/), while in some cases a special diacritic called *halant* (◌्) is added below the character (eg. ट् + क → ट्क /ʈk/). This will arise when the consonant is part of a conjunct consonant. In all such cases a cost of .5 is assumed.

5. **Ligatures and composite characters**: Unlike the above cases of conjunct consonants, lack of vowel on one of the consonant can some times cause the ligature to take a complex form (for example, /t/ + /rə/ = त् + र → त्र; /r/ + /tə/ = र् + त → र्त; /k/ + /tə/ = क् + त → क्त). In all such cases a cost of 1 is posited. A cost of 1 is also posited for composite characters such as क्ष /kʂə/ and ज्ञ /gyə̃/.

In addition, in cases such as कार्निवाल /kɑrnıval/ where /r/ has been displaced further due to the intervening ि◌ /ı/, the cost incorporates the distance of displacement.

The mean word complexity in the Hindi text was 0.46 (minimum: 0, maximum: 5.5).

Table A3
Word complexity cost

Factor	Cost	Additional displacement cost
Vowel appears to the left of a consonant	1	Yes
Diacritic above a consonant	0.5	No
Diacritic below a consonant	0.5	No
Consonant (no inherent vowel)	0.5	No
Ligatures	1	Yes

Appendix B
The sentences used in the study

The complete set of items are available from: http://web.iitd.ernet.in/∼samar/data/hindi-data.tar.