

Effect of jumbling the order of letters in a word on reading ability for Indian languages: An eye-tracking study

Bharat Ram Ambati (ambati@students.iiit.ac.in)

Language Technologies Research Centre, IIIT-Hyderabad, India.

Ganeshwar Rao Dulam (grdulam@students.iiit.ac.in)

Language Technologies Research Centre, IIIT-Hyderabad, India.

Samar Husain (samar@iiit.ac.in)

Language Technologies Research Centre, IIIT-Hyderabad, India.

Bipin Indurkha (bipin@iiit.ac.in)

Cognitive Science Lab, IIIT-Hyderabad, India.

Abstract

The paper describes various reading experiments to investigate human reading patterns for Hindi and Telugu. Experiments are based on different types of word jumbling. Results from these experiments bring out some interesting observations which are related to orthography, morphology, syntax, semantics, and discourse. These patterns were noticed across majority of the subjects. Some of these observations are unique to Hindi and Telugu.

Keywords: Reading, eye-tracking, Indian languages, Reading model.

Introduction

Many empirical studies have been done to understand people's reading patterns (Rawlinson, 1976; Healy, 1976; McCusker, 1981; Shillcock, 2000). For example, it has been found that when the text consists of familiar words, jumbling the middle letters of words in a text does not impair reading (Perea & Lupker, 2003a; Perea & Lupker, 2003b; Larson, 2007). A few reading models have been proposed (Reichle et al. 1999; Rayner, 1998; Larson, 2007) which among other things try to account for such reading patterns. However, these models are based on data for English (Clifton et al., 2007), which is not a phonetic language. It would be interesting to see if these models apply to phonetic languages as well and, if not, how they may be modified or extended to include them as well. Towards this goal, we have been designing and conducting eye-tracking experiments to glean reading patterns for Hindi and Telugu, two Indian languages that are phonetic. In this paper, we report on the results of some preliminary experiments.

Experiments Overview

For stimulus, we extracted Hindi¹ and Telugu² text from their respective Wikipedia. The content was a brief

¹ Hindi is a verb final Indo-Aryan language with free word order and a rich case marking system. It is one of the official languages of India, and is spoken by ~422 million people.

description of Hyderabad (State capital of Andhra Pradesh, India). Extracted text was manually corrected for spelling and grammatical errors. Table 1 gives some basic characteristics of the text for each language.

Table 1: General Statistics.

	Hindi	Telugu
Total no. of Words	169	103
Average no. of syllables per word	2.73	3.95
Average no. of characters per word	4.92	7.81

A total of 60 subjects participated in the experiments; 35 subjects for Telugu and 25 for Hindi. All the subjects were undergraduate and graduate students between 20–27 years of age, with a mean age of 22 years. For Telugu all the subjects were native speakers of the language. In the case of Hindi some of the subjects were not native speakers but spoke it as their second-language and knew the script well.

A total of 5 experiments were conducted for each language. For each of these experiments the data set used is described in Table 2.

Table 2: Data sets description. (C=consonant, V=vowel)

Data	Description
Set1	Original text. Unchanged. $C_1V_1C_2V_2C_3V_3C_{41}C_{42}V_4C_5V_5$
Set2	Jumbling of vowels $C_1V_1C_2V_3C_3V_2C_{41}C_{42}V_5C_5V_4$
Set3	Jumbling of consonants $C_1V_1C_3V_2C_2V_3C_5V_4C_{41}C_{42}V_5$
Set4	Jumbling of syllables (short distance)

² Telugu is a Dravidian agglutinative language. It is one of the official languages of India, and is spoken by ~74 million people.

	$C_1V_1C_3V_3C_{41}C_{42}V_4C_2V_2C_5V_5$
Set5	Jumbling of syllables (long distance) $C_1V_1C_{41}C_{42}V_4C_3V_3C_2V_2C_5V_5$

As is clear from table 2, data sets 2-5 are modifications of the original text (set 1). Different types of jumbling were explored. In set 2 and 3 we fix the leftmost syllable³, while jumbling the vowels and consonants respectively. Note that in both these sets, jumbling is done over short distance, i.e. the displacement never exceeds 2 units. In set4 and set5 the leftmost and the rightmost syllables are fixed and the intervening syllables are jumbled over short (≤ 2) and long (> 2) distance respectively.

Each subject was given one set to read while the eye-tracker was used to track his/her eye-movements. We also recorded the subject's voice for each reading session. For Telugu each set was read by 7 subjects while for Hindi 5 subjects read each set.

Base Statistics

The average reading time for Hindi set1 was 1.3 words/sec (wps). It was 1.1 wps for Telugu. The fixation point for both Hindi and Telugu was slightly left to the centre. It was observed that if the word was recognized (correctly or incorrectly) then the subject jumped to the next word without fixating on the remaining word segments (Rayner, 1979).

Experiments

In this section we present the results for each of the five experiments in turn, and discuss their implications. Table 3 lists the average reading time and average error rate for all the experiments.

Experiment 1

Set1 being the original text, the average reading time and error rates are the least among all the sets for both Hindi and Telugu. Subjects made very minor mistakes while reading the text. It is interesting to note that most of the mistakes were grammatically and contextually valid, though the lexical items read were absent in the text. In both Hindi and Telugu these errors were mostly due to adpositions. In Telugu the adpositions appear as suffixes where as in Hindi they are post-positions.

Experiment 2

In set2, only vowels were jumbled. The predictions were relatively easy as the subjects mostly guessed the appropriate vowels based on the unchanged consonants. A direct correlation was observed between the word length and the correct prediction. This is because a long word will tend to have more consonants that will tend to restrict the

search space. The opposite holds true for short-length words, for which the error rate is more than for long words (Figure 1, 2).

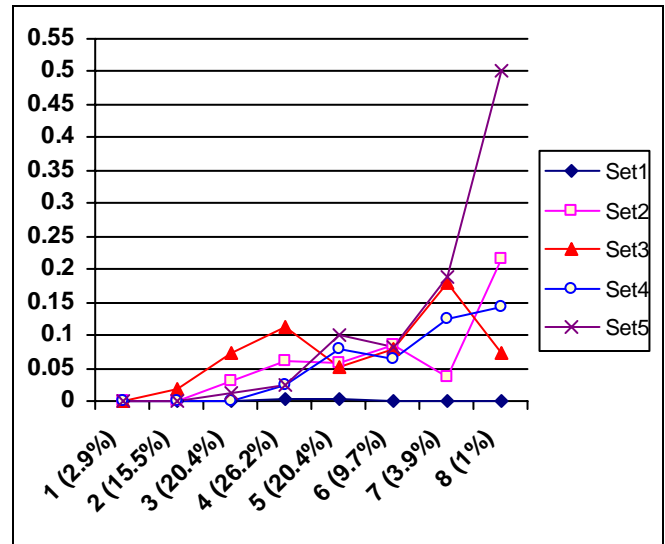


Figure 1: Telugu error rate w.r.t. syllable length

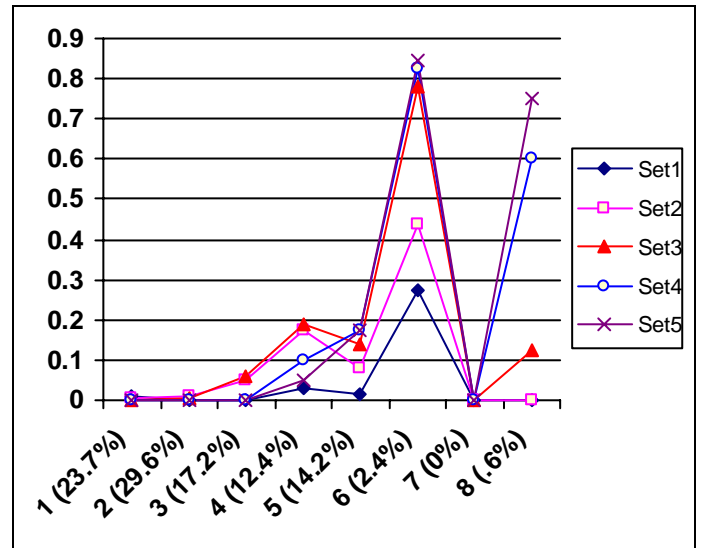


Figure 2: Hindi error rate w.r.t. syllable length

Experiment 3

In set3, only the consonants were jumbled. Understandably, the correct predictions become scarce as the information provided by the word pattern does not restrict the search space. For both Hindi and Telugu, this experiment showed highest error rate and highest reading time.

Experiment 4

As the first and last syllables are fixed and the jumbling is done over short distance, the prediction was relatively easier

³ A syllable or more specifically an open syllable follows the following pattern: C*V, where C=consonant, *=zero or more occurrence of the preceding element, and V=vowel.

than set2 and set3. It is easy to see why this is so. When compared to an individual unit, such as a vowel or a consonant, a syllable would have more information encoded in it. In set2 and set3, only part of the syllable was available. In such a case the syllable should be guessed first from the partial information and only later the word can be guessed. In set4, on the other hand, the jumbling is at the syllable level and that too at short distance. One only needs to guess the word directly here. This seems a plausible explanation to account for the subjects taking less time to read the text and making few mistakes for both Hindi and Telugu.

Experiment 5

Set5 is similar to set4, but has long distance jumbling. When compared to set4, the syllables are moved farther from their original positions. This results in a higher reading time and a higher error rate than set4. Though the displacement is long, an entire syllable is jumbled which makes more information available to predict the word than set3. Thus the error rate and reading time are lower than set3. This can be seen in Table 3.

Table 3: Average reading time and average error percentage.

	Hindi		Telugu	
	Time (Seconds)	Error %	Time (Seconds)	Error %
Set1	129.7	0.947	94.2	0.161
Set2	151.3	5.030	122.4	4.577
Set3	186.4	6.879	140.6	7.282
Set4	143.8	4.882	103.6	3.467
Set5	160.0	5.444	109.4	4.935

Factors affecting reading efficiency

In general, the reading efficiency was high for all the experiments. This means that the subjects are making many predictions which turn out to be correct. In this section we try to list out some of the factors that we think affect the subject's ability to predict correctly. These factors would thereby also affect reading efficiency with respect to speed and error rate.

Corpus type

Table 4 shows the difference in the reading time of the subjects when we vary the difficulty level of the text. It also shows the error rate. All the figures have been normalized. As expected, it is easier to read an easy text and much more difficult to read a text which has technical/domain specific words. Note that this in a way also reflects the subject's vocabulary for a given language. A subject with good vocabulary will not find a hard text difficult. This observation is consistent with that of Williams & Morris (2004) and Ashby et al. (2005).

Table 4: Corpus type

		Easy	Moderate	Hard
Hindi	No. of words	83	169	89
	No. of jumbled words	6	50	23
	Reading Time (wps)	2.31	1.18	1.08
	Error Percentage	0.40%	4.88%	4.87%
Telugu	No. of words	39	103	53
	No. of jumbled words	13	61	23
	Reading Time (wps)	1.45	1.06	1.01
	Error Percentage	5.97%	4.94%	10.68%

Context

Context plays a very crucial role in the subject's ability to make correct predictions. The following context types were used consistently by most of the subjects.

Individual words Predictions at the word level are a common strategy employed by the subjects. Incomplete segments of a word are used to predict the entire word. In the examples below the bold segments were used as cues to predict the correct word.

Hindi: महानगर 'metro' was guessed from the jumbled word मरनगर

Telugu: విశ్వ-విద్యాలయం 'University' was guessed correctly from the jumbled word వి-విశ్వ-లద్యాలయం

Immediate context In a large number of cases, particularly for function elements such as post-positions, the preceding fixation point is enough to read these elements. This means subjects usually skip these words during the saccades. Other cases where local context becomes pertinent are compounding and grouping of words. Some such examples are:

Telugu: In సికాలీసు వ్యాలీ 'Silicon Valley' the bold text could not be read in the first gaze, but was identified correctly after reading the second part of the compound 'Silicon valley'.

Hindi: The word आर्कपण was identified correctly only after the next word was processed correctly. The next word was चारमीनार (a historical monument).

Sentence level Similar to the previous case, we observed a large number of instances when the auxiliaries of the verbal unit were skipped. Along with this, many a times the

agreement suffixes are also guessed. This is not surprising as in both Telugu and Hindi the verb agrees with the subject. Hence, once the properties of the subject are known one can guess the verbal forms. A similar study has been done by Deutsch and Bentin (2001).

Previous discourse We observed that some of the subjects made considerable use of the previous discourse while making the prediction. Things such as sentence style, topic, theme, etc. from the previous 2-3 sentences were used to make prediction. In one such case, in Telugu a sentence ended with a verb పొందినాయి ‘attain-PAST’ which is an old Telugu form. The next sentence ended with a verb నిర్మించారు ‘build-PAST’ which is a new form. But some subjects guessed this verb in its old form నిర్మించిరి ‘build-PAST’. For Hindi, in some instances the subject carried forward the previous sentence grammatical subject pattern. For example, they expected the present sentence subject to be in oblique case, since the previous two sentences’ subjects were in oblique case.

All four contexts can be used for either *predicting* or for *correcting* previous predictions. For example:

Individual word level prediction: A jumbled word like आर्कषण read as आरकषण ‘reservation’.

Immediate context level correction: The subject later backtracked and correctly read it as आकर्षण ‘attraction’ after reading the next word चारमीनार (a historical monument).

Indian language specific observations

Indian languages like Telugu and Hindi diverge from English at various levels. One such level is orthographic. The Devanagari script and the Telugu script, because of their characteristics, allow for different types of jumbling. In the experiments above we jumbled only vowels (Set2), only pure consonants (Set3), syllables (4 and 5). We tried to jumble the words at the phonemic level but we had to discard it because it made just no sense (Set0). We saw in Table 3 that short-distance syllable jumbling lends itself for easier reading than other kinds of jumbling. For very long words the subjects tend to guess at the morpheme level and try out various combinations, one such case in Hindi was नगरमहापालिका ‘PWD’ which was guessed as महा-नगर-पालिका, महा-नगर-निगम and नगर-महा-पालिका. We plan to explore this aspect further in our future research.

Table 4: Language-specific examples for different test sets.

	<i>Telugu</i>	<i>Hindi</i>
<i>Original/set1</i>	హైదరాబాదు	हैदराबाद
<i>Set2</i>	హైదారబుదా	हैदारबदा
<i>Set3</i>	హైరదాదాబు	हैरदादाब
<i>Set4</i>	హైరాబాదదు	हैराबादद
<i>Set5</i>	హైబారాదదు	हैबारादद
<i>Set0</i>	హైరలబ్బాళఉ	हैरआब्दआ

Due to the agglutinative nature of Telugu, the average word length is more than that of English. This is reflected in the average time taken to read a word, which is greater than that of English. For English this is 3 wps, and for Telugu it’s 1.3 wps. Also relevant is the basic syntactic structure of Telugu and Hindi. Unlike English these are SOV⁴ languages. In both these languages the verb agrees (in gender, number and person) with the subject of the sentence, this information appears on the verb as suffixes or auxiliaries. We must note here that the distance between the verb and the subject will be greater than English. We observed earlier that this information is generally skipped while reading. This means that in spite of this long distance the subjects are able to retain this information when they encounter the main verb of the sentence. We still need to explore the upper limit of this distance to see how it is related to the subject’s retention memory.

Both Telugu and Hindi make extensive use of post-positions and suffixes to mark the grammatical/thematic roles of nouns in a sentence. Like the agreement features these function words are also frequently skipped by the subjects during the saccades. What is interesting about this characteristic is that some of these post-positions/suffixes help in predicting the lexical choice and morphological properties of the verb.

Discussion

It is clear from Table 3 that even the worst error rate for both the language is not very high. Almost all the subjects were able to guess most of the words even when they were jumbled. Although the time taken for reading it varied based on the kind of jumbling. Once a word has been guessed (correctly or incorrectly) the subjects were faster at making their prediction in case of repetition of that word (Rayner et al., 1995). Interestingly, if the same word ended up being jumbled in different ways in the same text, the subjects mostly were able to identify it as the same word. Also, the subjects usually stick to their first guess, if that guess seems to fit in the present context. Only if their first prediction

⁴ Subject-Object-Verb

does not seem right (based on the type of the context described earlier) do they make the second guess.

While reading the text, the subjects almost consistently skip the function words, i.e. do not focus on it. The reading also involves different kinds of operations on words, namely merging, addition, substitution. Below we give some examples for each of these operations:

Telugu, *addition*: అంతేకాదు ‘not only’ when jumbled as

‘అంతాకుదే’ read as ‘అంతకుముందే’ ‘before hand’. The ‘ముం’ was added.

Telugu, *merging*: షా పదిహేను ‘Shah 15’ jumbled as షా

పదహారుని read as షాజహాను ‘Shah Jahan’.

Hindi, *substitution*: , ‘comma’ read as और ‘and’.

Hindi, *merging*: जाना जाता ‘is believed’ read as जानता ‘knows’.

Tentatively, the best model which fits the statistics and the observations seems to be *E-Z Reader 3* (Reichle et al., 1999). However, we need to further flesh out the results and do some more experiments to say this for sure.

Conclusion and Future work

In this work we tried out various reading experiments where texts from two Indian languages were jumbled. The orthography of these languages allowed us to try out different types of jumbling. It was observed that in spite of the jumbling, subjects consistently read the text correctly, although the time taken by them to read different texts varied. Several interesting results came to the fore, and some of them were because of the characteristics of Indian languages. We plan to perform further experiments to elaborate and consolidate our observations and results in more detail. Tentatively, the results seem to fit E-Z Reader 3 model.

References

- Ashby, J., Rayner, K., & Clifton, C.J. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology*, 58A, 1065-1086.
- Clifton, C., Jr., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In Van Gompel, M. Fisher, W. Murray, and R. L. Hill (Eds.), *Eye movement research: A window on mind and brain*. Oxford: Elsevier Ltd. pp. 341-372
- Deutsch, A., & Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew: Evidence from ERPs and eye movement. *Journal of memory and Language*, 45, 200-224.
- Healy, A. F. (1976). Detection errors on the word *The*: Evidence for reading units larger than letters. *Journal of Experimental Psychology: Human Perception & Performance*, 2, 235-242.

- Larson, K. (2007). The Science of Word Recognition or how I learned to stop worrying and love the bouma. <http://www.microsoft.com/typography/ctfonts/WordRecognition.aspx>
- McCusker, L. X., Gough, P. B., Bias, R. G. (1981). Word recognition inside out and outside in. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3), 538-551.
- Perea, M., & Lupker, S. J. (2003a). Transposed-letter confusability effects in masked form priming. In S. Kinoshita and S. J. Lupker (Eds.), *Masked priming: State of the art* (pp. 97-120). Hove, UK: Psychology Press.
- Perea, M., & Lupker, S. J. (2003b). *Does judge activate COURT? Transposed-letter confusability effects in masked associative priming*. Memory and Cognition.
- Rawlinson, Graham. (1976). *The Significance of Letter Position in Word Recognition*. PhD Thesis, Nottingham University.
- Rayner, K. (1979). Eye guidance in reading: fixation locations within words, *Perception*, 8(1) 21 – 30.
- Rayner, K., Raney, G., & Pollatsek, A. (1995). Eye movements and discourse processing. In R. F. Lorch & E. J. O’Brien (Eds.), *Sources of coherence in reading* (pp. 9–36). Hillsdale, NJ: Erlbaum.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372–422.
- Reichle, ED, Rayner, K., & Pollatsek, A. (1999). Toward a model of eye movement control in reading, *Psychological Review*, 105, 125-157.
- Shillcock, R., Ellison, T.M. & Monaghan, P. (2000). Eye-fixation behaviour, lexical storage and visual word recognition in a split processing model. *Psychological Review* 107, 824-851.
- Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, 16, 312–339.