A Metric for Lexical Complexity in Malayalam

Richard Shallam (imrichsham@gmail.com)
Independent Researcher

Ashwini Vaidya (avaidya@hss.iitd.ac.in)
IIT Delhi

Abstract

This paper proposes a metric to quantify lexical complexity in Malayalam. The metric utilizes word frequency, orthography and morphology as the three factors affecting visual word recognition in Malayalam. Malayalam differs from other Indian languages due to its agglutinative morphology and orthography, which are incorporated into our model. The predictions made by our model are then evaluated against reaction times in a lexical decision task. We find that reaction times are predicted by frequency, morphological complexity and script complexity. We also explore the interactions between morphological complexity with frequency and script in our results. To the best of our knowledge, this is the first study on lexical complexity in Malayalam.

Keywords: lexical processing, visual word recognition, lexical complexity, Dravidian languages

1 Introduction

The task of visual word recognition is related to language processing at the level of a word/lexical item. A word can be analyzed at several linguistic levels, and the word recognition task helps us understand the role of these levels in relation to processing, memory and attention. In psycholinguistics, previous work on this topic focuses on understanding the individual variables that affect the lexical processing of words. If we can quantify the influence of variables ranging from orthographic features to semantic factors on the cognitive processing of words, it would help us in understanding the critical factors underlying visual word recognition (and pattern recognition, more generally). The resulting model of word recognition can be evaluated against human judgements.

Models of word recognition are especially relevant for eye-tracking studies, where they have been extensively explored (Rayner and Duffy, 1986). Word recognition models have also been used to understand reading disabilities such as phonological and surface dyslexia (Balota et al., 2006). For these studies, it is crucial to tease apart the effect of various factors that affect the task of reading. Previous research has shown that the eye gaze duration is affected by frequency, orthography, morphology and phonology, among others. Apart from these studies, an understanding of lexical complexity is also an interesting topic for study on its own.

In this paper, we explore the case of Malayalam and in particular examine three factors that could predict word complexity in the language: frequency, orthography and morphology. The role of variables that determine word recognition in Malayalam has not been explored, as it has been for Hindi (Husain et al., 2015; Verma et al., 2018). Quantifying these factors in a model of lexical complexity can help us in developing norms that are useful in areas such as reading studies and word generation for lexical decision tasks. Further, this would contribute towards cross-linguistic comparison of these factors from a different language family. To the best of our knowledge, this is the first work that examines lexical complexity in Malayalam.

2 Lexical Complexity

The task of visual word recognition involves the cognitive processing of visual information and comparing it with a particular internal

D M Sharma, P Bhattacharyya and R Sangal. Proc. of the 16th Intl. Conference on Natural Language Processing, pages 178–183
Hyderabad, India, December 2019. ©2019 NLP Association of India (NLPAI)

mental representation of a word. This representation itself may be at the graphemic, phonemic, morphemic and lexical semantic level, all of which have been shown to affect word recognition (Balota et al., 2006). In the sections that follow, we describe the three factors that are included in our study.

2.1 Word Frequency

The effect of word frequency is robust and has been well studied across word recognition tasks (Balota et al., 2006). High frequency words tend to be recognized faster than low frequency words. In eye tracking studies high frequency words have lower gaze duration and fixation measures. We would expect that frequency would have a similar effect on the Malayalam data, where high frequency would contribute towards a lower lexical complexity.

2.2 Morphology

A word may be composed of a single morpheme e.g. *boy* or more than one e.g. *funnily: funny+ ly.* The role of morphology in word recognition is at a sub-lexical level. Morphology as a measure is particularly relevant for an agglutinative language such as Malayalam, which also exhibits productive word compounding e.g. Just the word മരം (mara) "tree" has a number of morphological forms such as

മരത്തിൽ	(marattil) - in the tree
മരത്തിന്റെ	(marattinṟe) - of the tree
മരങ്ങൾക്കിടയിലൂടെ	(maraṅṅaḷkkiṭayilūṭe) - through the trees
മരക്കൊമ്പുകൾ	(marakkeāmpukaḷ) - tree branches

Early studies that looked at the effect of morphology on lexical access have suggested that polymorphemic words (i.e. words consisting of more than one morpheme) are decomposed into their component parts during online processing. This process would find the root first (e.g. *funny* and on finding it, proceed to search stored affix-stem combinations till *funnily* is retrieved (Taft and Forster, 1975). In a morphologically-rich language such as Malayalam, we would expect that this would be an important factor in lexical processing.

2.3 Orthography

The visual processing of words involves processing at the orthographic level as well. This implies that the writing system of various languages will influence recognition. A writing system–whether alpha-syllabic, logographic or alphabetic has been shown to influence reading times (Katz and Frost, 1992). Sub-lexical properties such as letter features and their interactions with the words themselves can also influence word complexity, which needs to be accounted for in the model.

3 Method

In order to compute the lexical complexity metric, token frequency, morphology and orthography were included as our variables. Below, the methods for computing the values for each of these variables are discussed.

3.1 Corpus

In order to compute our metric for Malayalam, we first obtained a corpus from the Leipzig Corpora Collection containing 300,000 sentences from Malayalam Wikipedia articles and 100,000 sentences from Malayalam news crawl (Goldhahn et al., 2012). The corpus was then preprocessed by removing punctuation and special characters, and then tokenized using whitespace. The text was also normalized to remove inconsistencies in spelling using the Indic NLP Library[1] and this resulted in 4,711,219 tokens and 762,858 unique types.

3.2 Word Frequency Metric

The corpus was used to collect counts for each word and then scaled them between 0 and 1, which was then inverted such that the most frequent tokens have a value closer to 0 and the less frequent tokens will have a value approaching 1. This score indicated the relative frequency of each word in this corpus, and the idea that highly frequent words are much easier to process than those that have lower frequency.

3.3 Morphology Metric

Our morphology metric required us to obtain information about the root and the morpho-

[1] https://anoopkunchukuttan.github.io/indic_nlp_library/

logical affixes for a given word. Given the rich morphology and compounding processes in the language, we had to make use of a two-step process to compute our scores.

First, *SandhiSplitter* (Devadath et al., 2014) was used to split tokens that are compound words into their constituent component words. For example, consider the compound word കാരണമായിരിക്കണം (kAraNamAyirikkaNaM)

കാരണമായിരിക്കണം ⇒ കാരണം + ആയിരിക്കണം

kāraṇamāyirikkaṇaṁ ⇒ kāraṇaṁ + āyirikkaṇaṁ

“must be the reason” ⇒ “reason” + “must be”

As a second step, these results were passed through IndicStemmer[2], a rule-based stemmer for Malayalam, which further decomposed the words into stems and affixes. As an example, the word ലേഖനങ്ങളുടെ (lēkhanaṅṅaḷuṭe) meaning “Of articles”. is decomposed into the stem ലേഖനം (lēkhanaṁ) meaning article with the suffix -ങ്ങൾ (ṅṅal) indicating plural and --ുടെ (uṭe) indicating the Genitive case. In our metric we only considered suffixes as in Malayalam usually contains always suffixes being added to the end of the stem.

After this two-step process, we are able to obtain the stems and suffixes for a given word.

Morpheme Count

By simply summing the number of stems and suffixes, the total number of morphemes contained in each word is computed. For example, the word സമ്പത്സമൃദ്ധിയും (sampatsamrd’dhiyuṁ) meaning “prosperity” is a compound word split into constituent words സമ്പത്ത് (sampatt) meaning “richness” and സമൃദ്ധിയും (samrd’dhiyuṁ) meaning “and plentiful”. സമൃദ്ധിയും (samrd’dhiyuṁ) is further stemmed to stem word സമൃദ്ധി (samrd’dhi) meaning ”plentiful” and suffix -ും (uṁ) meaning ”-and”. സമ്പത്ത് (sampatt) is a root word. Thus, the number of morphemes in this case is three, counting the two stems and one suffix.

Based on this pre-processing, we then calculate the total number of morphemes for each whole word and then scale this number between 0 and 1 to give a morpheme score. We note that there could be several different ways to compute the morpheme score, as affixes themselves are not all alike. In this preliminary study, it was not immediately apparent how the differing costs for various affixes could be calculated. Additionally, fine-grained information regarding the morphological properties of the affixes (e.g. whether they were inflectional or derivational) was not easily obtained with existing tools and resources. In future work, we plan to explore this possibility by enhancing the morphological analyzer’s output.

3.4 Orthography Metric

Malayalam is an alphasyllabic writing system that has its source in the Vatteluttu alphabet from the 9th century. Its modern alphabets have been borrowed from the Grantha alphabet. It consists of 15 vowels and 36 consonant letters.

We devised a script score based on complexity of the script in the following three ways:-

Mismatch in Spoken and Visual Order

In the alpha-syllabic script of Malayalam, vowels may either appear as letters at the beginning of a word or as diacritics. Consonants themselves are understood to have an inherent schwa, which is not separately represented. The diacritics will appear either left or right of the consonant it modifies. If it appears to the left, there will be a discrepancy in the phonemic and the orthographic order, as the vowel will always be pronounced after the consonant, but read before the consonant actually appear in the text. For example:

ക +െ◌ = കെ

ka + .e = ke

Here the vowel violates the order in which it is spoken. Similarly: ക +േ◌ = കേ (ka + ē = kē), as seen in കേൾക്കുക (kēḷkkuka) meaning “hear”. Such inconsistencies in spoken and visual order have been shown to incur a cost in Hindi word recognition (which is also an alpha-syllabic script) (Vaid and Gupta, 2002).

In order to capture the lexical processing cost for such a discrepancy, we give a penalty of 1 every time it occurs in the word.

[2] https://github.com/libindic/indicstemmer

Diacritic Appearing Above or Below

In Malayalam, the diacritic may also appear above or below a consonant. In such a case, we we give a penalty of 0.5 to the word. For example the symbol ◌് also known as *virama* is used to replace the inherent schwa sound of consonants with ŭ. As in ക + ◌് = ക് (ka + virama = ku)

Ligatures and Consonant Clusters

A penalty of one is assigned for every two letters that form a composite glyph. For example: മന്ത്രി (mantri) = മന് + ത്രി (man + tri) where the new composite glyph is ന്ത്ര (ntra).

With the above complexity rules in place, the total penalty cost for each whole word is calculated. Then the total penalty for each word is scaled linearly to between 0 and 1 to give us an orthographic score.

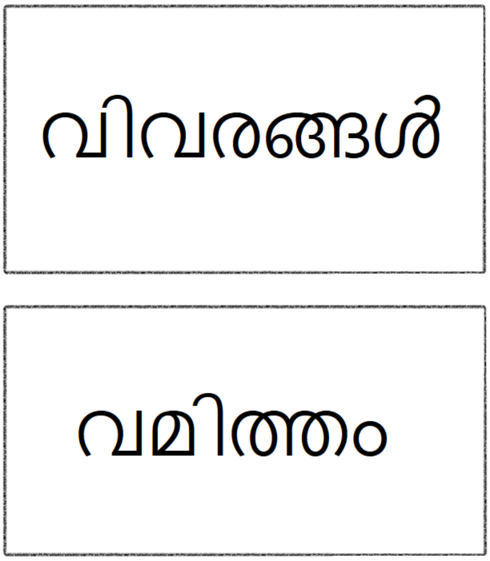
3.5 Evaluation of the Complexity Metric

In order to evaluate our lexical complexity metric, we used a lexical decision task paradigm to collect reaction times for a sample of Malayalam words. More complex words would result in longer reaction times, and vice versa. This would help us evaluate whether our lexical complexity model could predict reaction times for the given set of words.

We used a well-understood experimental paradigm in the form of a lexical decision task. In such a setup, a participant will see a word stimuli on a screen which they have to classify as either a word or a non-word using a button press. The response time (RT) is calculated from the point the word appears on the screen to the point where the participant presses the response button.

Materials

Our task consisted of a balanced set of 50 Malayalam words and 50 pseudowords. Pseudowords follow the phonotactics of the language, but have no lexical meaning (i.e. are not legitimate words). In order to select words for the task, two sets of 25 words were randomly sampled from the unique tokens obtained from the Leipzig Corpus. The first set was randomly sampled from words with a frequency score between the range of 0.1 to 0.4 to obtain high frequency words as calculated by the metric. The second set was chosen similarly but with frequency score between the range of 0.7 to 0.9 to yield low frequency words. If the sampled word turned out to be an English word written in Malayalam or happens to be a proper noun, it was replaced with another until both sets had 25 words each.



വിവരങ്ങൾ

വമിത്തം

Figure 1: Stimuli word shown for 2500ms. The first word is a proper Malayalam word (“vivaraṅṅaḷ” meaning “information”) hence the correct response is to press the ‘a’ key. The second word is non-word (vamittaṁ) and therefore, the correct response is to press ‘l’ key.

The pseudowords were constructed in keeping with the phonotactics of Malayalam. Both the pseudowords and the valid words were constrained in length between 6 and 14 characters. Note that we do not take into consideration the reaction times for the pseudowords; they are simply distractors for the participants.

Participants

Participants included 38 students from S.N. College, Kerala, who volunteered for the study. Participants included 20 females and 18 males between the ages of 18 and 23 (mean age of 19.7). All participants were native speakers of Malayalam and had formal education in Malayalam upto grade 10.

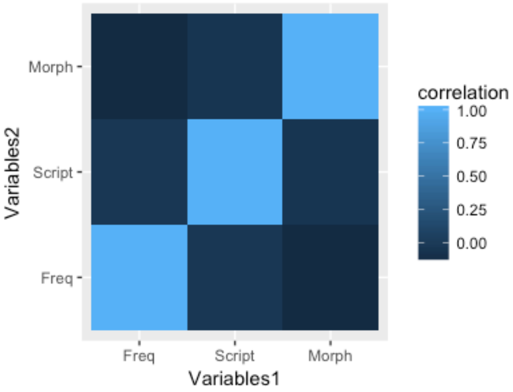


Figure 2: Heat plot showing correlation between the three variables in our test data

Procedure

Participants were tested individually on a computer running the lexical decision task on the JsPsych stimulus presentation software (De Leeuw, 2015). Each participant was asked to press either the 'a' key or the 'l' key for word and non-word respectively. The order of words and pseudowords was randomized for each participant. Participants were instructed to read the word presented and respond with the appropriate button press. Each trial consisted of a word that was presented for 2500ms. A fixation cross was placed in the center for 1600ms between each trial. The first 10 trials were practice trials from a word set different from the study. This enabled participants to get familiarized with the task.

4 Results

The trials belonging to those who scored below 70% in word-non-word accuracy were excluded, which brought the number of participants to 35.

We fit a linear model using the `lm` function in R. Log reaction times were used with frequency, script and morph as the covariates. Figure 2 shows that the three variables are not highly correlated in our test set.

Table 1 shows the results of the regression analysis. The main inference we can draw from the result is that the variables Script, Morphology and Frequency have a significant effect (all p-values < 0.05) on (reaction times) RTs, such that a high cost of script, morph and frequency leads to higher RTs.

In addition, the results also indicate a marginal interaction between Script and Morphology (p=0.06), such that an increase in the script complexity leads to larger increases in RTs for morphologically simpler words (Cost <0.9) compared to morphologically complex words (Cost >0.9) (see Figure 3). There is also a marginal interaction between Morphology and Frequency (p=0.08) such that an increase in the frequency cost leads to higher reaction times in morphologically complex words as compared to morphologically simpler words (see Figure 4).

	Estimate	***Std. Error***	***t-value***	***p-value***
(Intercept)	4.30	0.679	6.35	0
Script	9.157	3.76	2.43	**0.015 ***
Freq	2.87	0.96	2.97	**0.003 ****
Morph	1.91	0.71	2.67	**0.007 ****
Script:Freq	-3.171	5.77	-0.55	0.58
Script: Morph	-7.64	4.1	-1.873	**0.06 .**
Freq: Morph	-1.79	1.03	-1.743	**0.08 .**
Script:Freq:Morph	0.28	6.31	0.045	0.96

Table 1: Results for all three variables and their interactions. Script and Morphological Complexity as well as Frequency and Morphological Complexity show a significant interaction

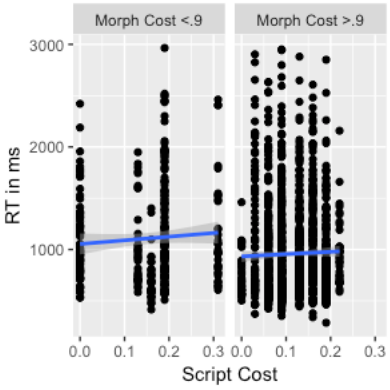


Figure 3: Interaction between Morphological Complexity and Script Complexity

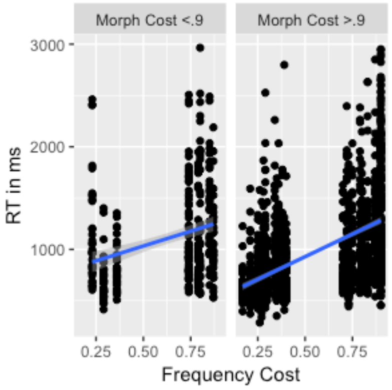


Figure 4: Interaction between Morphological Complexity and Frequency Cost. Note that a low Frequency Cost corresponds to a high Frequency Count for a word

5 Discussion

Our results replicate the robust effects of frequency on lexical processing in Malayalam. As frequency is a known predictor of reaction times, we expected to find a significant effect for frequency, but we particularly wanted to understand the effect of morphology and orthography on word recognition in Malayalam. Orthographic complexity as captured by diacritic placement and ligatures also has a significant effect on lexical processing. Similarly, we also find an effect for morphological complexity in terms of the number of morphemes in a word.

The interactions in our model point to an interesting relationship between high frequency words and morphological complexity. It appears that the effect of frequency cost becomes more pronounced in more complex words. In other words, low frequency words lead to higher reaction times particularly when they are morphologically complex. Perhaps this is because the cost of lexical decomposition is higher in these words. On the other hand, the effect size of script is weaker and becomes visible only when the word is morphologically simple. When the word is morphologically complex, this effect is not very apparent.

This work points to many interesting future avenues for exploring lexical complexity in an agglutinative language like Malayalam. Particularly, the effect of morphological complexity on factors like frequency need to be explored more thoroughly. In the future, we plan to carry out experiments with a larger set of items for the lexical decision task, as this was a preliminary study. We also plan to experiment with other measures of morphological complexity that take into account information about the type as well as the number of morphemes.

References

David A Balota, Melvin J Yap, and Michael J Cortese. 2006. Visual word recognition: The journey from features to meaning (a travel update). In *Handbook of Psycholinguistics*, pages 285–375. Elsevier.

Joshua R De Leeuw. 2015. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47(1):1–12.

VV Devadath, Litton J Kurisinkel, Dipti Misra Sharma, and Vasudeva Varma. 2014. A sandhi splitter for malayalam. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 156–161.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. 2015. Integration and prediction difficulty in hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, 8(2):1–12.

Leonard Katz and Ram Frost. 1992. The reading process is different for different orthographies: The orthographic depth hypothesis. In *Advances in psychology*, volume 94, pages 67–84. Elsevier.

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Marcus Taft and Kenneth I Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior*, 14(6):638–647.

Jyotsna Vaid and Ashum Gupta. 2002. Exploring word recognition in a semi-alphabetic script: The case of devanagari. *Brain and Language*, 81(1-3):679–690.

Ark Verma, V. Sikarwar, H. Yadav, J. Ranjith, and Pawan Kumar. 2018. Shabd: A psycholinguistics database for hindi words. In *Proceedings of ACCS 2018*.