Quantitative Analysis of the Morphological Complexity of Malayalam Language

Kavya Manohar[1,3][0000−0003−2402−5272], A. R. Jayan[2,3][0000−0001−7316−5808], and Rajeev Rajan[1,3]

[1] College of Engineering Trivandrum, Kerala, India
[2] Government Engineering College Palakkad, Kerala, India
[3] APJ Abdul Kalam Technological University, Kerala, India
sakhi.kavya@gmail.com

Abstract This paper presents a quantitative analysis on the morphological complexity of Malayalam language. Malayalam is a Dravidian language spoken in India, predominantly in the state of Kerala with about 38 million native speakers. Malayalam words undergo inflections, derivations and compounding leading to an infinitely extending lexicon. In this work, morphological complexity of Malayalam is quantitatively analysed on a text corpus containing 8 million words. The analysis is based on the parameters type-token growth rate (TTGR), type-token ratio (TTR) and moving average type-token ratio (MATTR). The values of the parameters obtained in the current study is compared to that of the values of other morphologically complex languages.

Keywords: Morphological Complexity · Types and Tokens · TTR · Malayalam Language

1 Introduction

Malayalam[4] is a language with complex word morphology. Malayalam words undergo inflections, derivations and compounding producing an infinite vocabulary [19]. As a language with high morphological complexity it has a large number of wordforms derived from a single root word (such as the English words *houses* and *housing*, which stem from the same root word *house*). Morphological complexity can be measured either in terms of the average number of grammatical features getting encoded into a word or in terms of the diversity of word forms occurring in the text corpus of a language. The former approach is called typological analysis and the latter one is called corpus based analysis of morphological complexity [5]. Morphological complexity of a language has its impact on applications like automatic speech recognition (ASR) where speech to text conversion depends largely on the underlying language model. A measure of the complexity is important for improving and adapting the existing methods of natural language processing (NLP) [10].

[4] https://en.wikipedia.org/wiki/Malayalam

This paper analyses the morphological complexity of Malayalam in terms of corpus based parameters namely, type-token growth rate (TTGR), type-token ratio (TTR) and moving average type-token ratio (MATTR). These parameters are formally defined in section 5. The study is conducted on a Malayalam text corpus of 8 million words.

2 Literature Review

Complexity of a natural language can be in terms of morphology, phonology and syntax [3]. Morphological level complexity of a language implies a large possibility of inflections (by grammatical tense, mood, aspect and case forms) and agglutinations (of different wordforms). The number of possible inflection points in a typical sentence, the number of inflectional categories, and the number of morpheme types are all morphological complexity indicators [4]. It requires a strict linguistic supervision to analyse each word in terms of its morpheme types to quantify complexity in this manner. Bentz et al. performed typological analysis of morphological complexity involving human expert judgement and compared it with corpus based analysis of morphological complexity and drew strong correlation between the two [5].

Covington et al. suggested the use of MATTR as a reliable measure of linguistic complexity independent of the total corpus length and suggested an efficient algorithm for computing MATTR [6]. Kettunen [13] compared corpus based parameters like TTR and MATTR with other methods of complexity measures as defined by Patrick Juola [12] and concluded both TTR and MATTR give a reliable approximation of the morphological complexity of languages. Ximena Gutierrez-Vasques et al. suggested estimating the morphological complexity of a language directly from the diverse wordforms over a corpus is relatively easy and reproducible way to quantify complexity without the strict need of linguistic annotated data [10].

3 Problem Statement

Malayalam has seven nominal case forms (nominative, accusative, dative, sociative, locative, instrumental and genitive), two nominal number forms (singular and plural) and three gender forms (masculine, feminine and neutral). These forms are indicated as suffixes to the nouns. Verbs in Malayalam get inflected based on tense (present, past and future), mood (imperative, compulsive, promissive, optative, abilitative, purposive, permissive, precative, irrealis, monitory, quotative, conditional and satisfactive), voice (active and passive) and aspect (habitual, iterative, perfect) [16,19]. The inflecting suffix forms vary depending on the final phonemes of the root words. Words agglutinate to form new words depending on the context [2]. Table 1 gives examples of a few complex word formation in Malayalam.

The productive word formation and morphological complexity of Malayalam are documented qualitatively in the domain of grammatical studies. However

Table 1: Complex morphological word formation in Malayalam

Malayalam Word	**Translation to English**	**Remark**
പെട്ടിയിൽ (peṭṭijil)	in the box	Nominal locative suffix to the word പെട്ടി (peṭṭi, box)
കുട്ടിയോട് (kuṭṭijo:ṭ)	to the child	Nominal sociative suffix to the word കുട്ടി (kuṭṭi, child)
ആനക്കുട്ടി (a:nakkuṭṭi)	baby elephant	Compound word formed by agglutination of nouns ആന (a:na, elephant) and കുട്ടി (kuṭṭi, baby)
ആനക്കുട്ടികളോട് (a:n̪akkuṭṭikaḷo:ṭ)	to the baby elephants	Nominal sociative suffix to the plural form of the compound word ആനക്കുട്ടി (a:nakkuṭṭi, baby elephant)
ഉണർന്നിരിക്കണ്ട (uṇarn̪n̪irikkaṇṭa)	do not stay awake	Negative imperative mood of the verb ഉണരുക (uṇaruka, be awake)
പാടിക്കൊണ്ടിരിക്കും (pa:ṭikkoṇṭirikkum)	will be singing	Future tense iterative aspect of the verb പാടുക (pa:ṭuka, to sing)

a quantitative study on the same is not yet available for Malayalam language. Adoption of general NLP solutions of high resource languages like English is not feasible in the setting of morphologically complex languages. A functional morphology anlayser, *mlmorph* addresses the morphological complexity of Malayalam applying grammatical rules over root word lexicon [19]. Quantification of linguistic complexity is important to adapt and improve various NLP applications like automatic speech recognition, parts of speech (POS) tagging and spell checking [9,14,17,18]. This study aims at quantifying the morphological complexity of Malayalam in terms of corpus parameters.

4 Material

This study is performed on Malayalam running text from Wikipedia articles. The Malayalam Wikipedia dump is curated and published by Swathanthra Malayalam Computing (SMC) as *SMC Corpus* [1]. It consists of 62302 articles. The Malayalam running text often has foreign words, punctuation and numerals present in it. The corpus is first cleaned up to eliminate non Malayalam content and punctuations. It is then unicode normalized [7]. The cleaned up corpus contained 8.14 million Malayalam words. The nature of the text is formal encyclopedic Malayalam.

5 Method

An element of the set of distinct wordforms in a running text is called a *type*. Every instance of a type in the running text is called a *token*. For example, in the sentence, *To be or not to be is the question*, there are 7 types and 9

tokens. The types *to* and *be* repeat two times each. The relationship between the count of types and tokens is an indicator of vocabulary richness, morphological complexity and information flow [10]. The type-token ratio (TTR) is a simple baseline measure of morphological complexity [13]. TTR is calculated by the formula defined in equation (1), where V is the count of types and N is the count of tokens.

$$TTR = \frac{V}{N} \tag{1}$$

The type count gets expanded due to productive morphology and higher values of TTR correspond to higher morphological complexity [5]. However TTR is affected by the token count, N [6]. Longer the corpus, it is more likely that the new tokens belong to the types that have occurred already. The value of TTR gets smaller with the increase in token count. Computing TTR over incrementally larger corpus can indicate how the TTR varies with the token count. In this study, TTR is computed with different token counts starting with 1000 and increasing upto the entire corpus size. This has enabled comparison of Malayalam with the morphological complexity of other languages whose TTR values are available in literature for different token counts.

The type-token growth rate (TTGR) curve is obtained by plotting the graph of token count vs. type count. It indicates how many new types appear with the increase in the token count. If the slope of the growth rate curve reduces and approaches a horizontal line, at a lower value of token count, it indicates a simple morphology [15]. For a morphologicaly complex language, the type count continues to grow with the token count [11].

The moving average type-token ratio (MATTR) computes the relationship between types and tokens that is independent of the text length. Its efficient implementation by Covington et al. has been used by Kettunen to compare the morphological complexity of different European languages [6,13]. The algorithm to compute MATTR is as follows [8]:

Algorithm 1: Computation of MATTR

Data: A text Corpus
Result: MATTR
N ← length of corpus;
L ← length of window (L<N);
start ← initial position of window ;
i = start ← index of window position;
while $i \leq (N - L + 1)$ **do**
  $V_i = type\ count\ in\ the\ window\ [i, i + L - 1]$;
  $TTR(i) = \frac{V_i}{L}$;
  $i = i + 1$;
end
$MATTR(L) = \frac{\sum_{i=1}^{N-L+1} TTR(i)}{N-L+1}$

The corpus with N tokens is divided into the overlapped subtexts of the same length, say L, the window length. Window moves forward one token at a time and TTR is computed for every window. MATTR is defined as the mean of the entire set of TTRs [6]. In this work L is chosen as 500, enabling comparison with other languages in the study by Kettunen, where the window length is 500 [13].

6 Result and Discussion

Counting the types and tokens on *SMC Corpus*, TTGR and TTR curves are plotted. Figure 1 shows the TTGR curve on the left and the TTR on the right. TTGR curve shows a steep rise initially. As the token count reaches 8 million, the type count is around 1.2 million. But the curve does not flatten even at that token count. This pattern is a common property of Dravidian languages as many unseen wordforms appear as the corpus size is increased [15]. TTR is very high at around 0.82 when the token count is 1000. TTR reduces to around 0.44 when the token count is 0.1 million and finally flattens to a value of 0.16 for the full corpus of 8 million tokens.

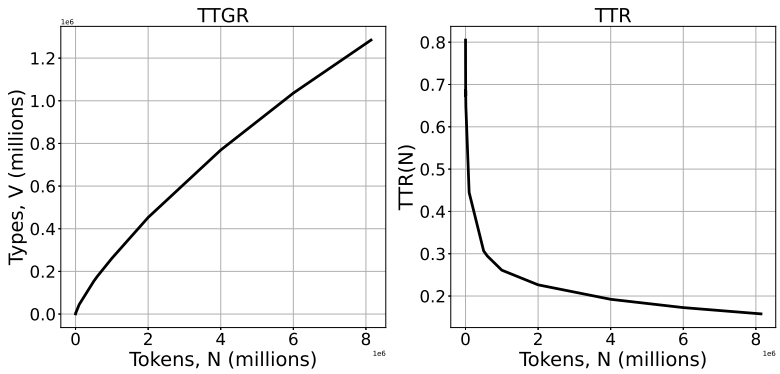
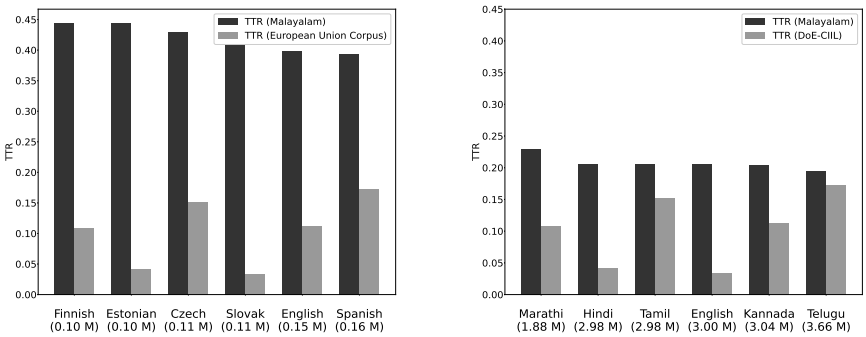


Figure 1: TTGR and TTR plot of Malayalam for *SMC Corpus* of Wikipedia text

To compare the TTR obtained for Malayalam with that of other languages, we have used the data reported for European languages by Kettunen and for Indian languages by Kumar et al. [13,15]. Figures 2a and 2b illustrates the comparison. Only those languages with the highest reported TTRs in the respective papers and English are used for comparison. The token size (in millions) used for computing TTRs used in the comparisons is indicated for each language. Malayalam clearly shows more morphological complexity than the European languages, Finnish, Estonian, Czech, Slovak, English and Spanish in terms of TTR values. Values of TTR obtained for Malayalam when compared with other In-

dian languages Marathi, Hindi, Tamil, Kannada and Telugu indicate a higher level of morphological complexity for Malayalam.



(a) (b)

Figure 2: Comparison of Malayalam TTR with that of *European Union Constitution Corpus* [13] and *DoE-CIIL Corpus* [15]

MATTR is computed with window length, L=500 over different segments of the *SMC corpus*. TTR values for the segments with window position index 1-1000, 5001-6000, 15001-16000 and 18001-19000 are plotted in Figure 3. These segments gave MATTR values 0.834, 0.839, 0.836 and 0.800 respectively. Computing MATTR with 0.1 million tokens of *SMC corpus* resulted in a value 0.806 for Malayalam. Kettunen has reported MATTR values on *European Union constitution corpus* with each language having a token count slightly above 0.1 million [13]. A comparative graph of the MATTR values reported by Kettunen with the values obtained for Malayalam is plotted in Figure 4. It clearly indicates a higher degree of morphological complexity for Malayalam in terms of MATTR on a formal text corpus. An equivalent comparison with other Indian languages could not be done due to non availability of reported studies.

7 Conclusion

In this paper we have reported a quantitative analysis of the morphological complexity of Malayalam language on a formal text corpus of 8 million words. The corpus based analysis has revealed high degrees of morphological complexity of Malayalam in terms of TTR and MATTR. It is important that this aspect of morphological complexity be considered while developing natural language processing applications like automatic speech recognition, spell checking and POS tagging for Malayalam. This involves preparing morpheme based language models and phonetic lexicons for ASR and performing a morphological analysis of words for POS tagging and spelling correction.

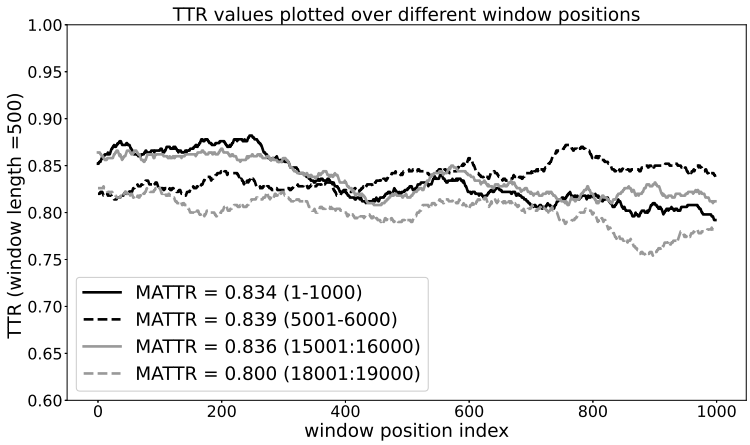


Figure 3: TTR plotted at different segments of the SMC corpus for 1000 window positions

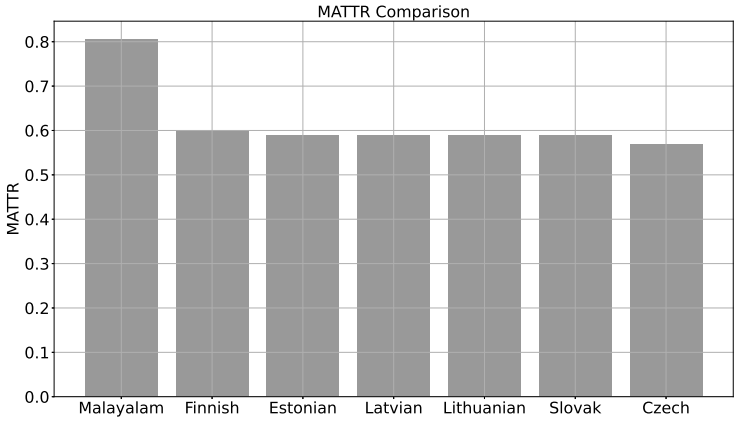


Figure 4: Comparison of MATTR values computed for Malayalam on *SMC Corpus* with that of *European Union Constitution Corpus* [13]

References

1. Malayalam Corpus. https://gitlab.com/smc/corpus (April 2020), by Swathanthra Malayalam Computing
2. Asher, R.E., Kumari, T.: Malayalam. Psychology Press (1997)
3. Baerman, M., Brown, D., Corbett, G.G.: Understanding and measuring morphological complexity. Oxford University Press, USA (2015)
4. Bane, M.: Quantifying and measuring morphological complexity. In: Proceedings of the 26th west coast conference on formal linguistics. pp. 69–76. Cascadilla Proceedings Project Somerville, MA (2008)
5. Bentz, C., Ruzsics, T., Koplenig, A., Samardzic, T.: A comparison between morphological complexity measures: Typological data vs. language corpora. In: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC). pp. 142–153 (2016)
6. Covington, M.A., McFall, J.D.: Cutting the gordian knot: The moving-average type–token ratio (mattr). Journal of quantitative linguistics **17**(2), 94–100 (2010)
7. Davis, M., Dürst, M.: Unicode normalization forms (2001)
8. Fidler, M., Cvrček, V.: Taming the Corpus: From Inflection and Lexis to Interpretation. Springer, 1 edn. (2018)
9. Georgiev, G., Zhikov, V., Osenova, P., Simov, K., Nakov, P.: Feature-rich part-of-speech tagging for morphologically complex languages: Application to bulgarian. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. p. 492–502. EACL '12, Association for Computational Linguistics, USA (2012)
10. Gutierrez-Vasques, X., Mijangos, V.: Comparing morphological complexity of Spanish, Otomi and Nahuatl. In: Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing. pp. 30–37. Association for Computational Linguistics, Santa Fe, New-Mexico (Aug 2018), https://www.aclweb.org/anthology/W18-4604
11. Htay, H.H., Kumar, G.B., Murthy, K.N.: Statistical analyses of myanmar corpora. Department of Computer and Information Sciences, University of Hyderabad pp. 1–15 (2007)
12. Juola, P.: Measuring linguistic complexity: The morphological tier. Journal of Quantitative Linguistics **5**(3), 206–213 (1998)
13. Kettunen, K.: Can type-token ratio be used to show morphological complexity of languages? Journal of Quantitative Linguistics **21**(3), 223–245 (2014)
14. Kipyatkova, I., Karpov, A.: Study of morphological factors of factored language models for russian asr. In: International Conference on Speech and Computer. pp. 451–458. Springer (2014)
15. Kumar, G.B., Murthy, K.N., Chaudhuri, B.: Statistical analyses of telugu text corpora. IJDL. International journal of Dravidian linguistics **36**(2), 71–99 (2007)
16. Nair, R.S.S.: : A grammar of malayalam. Language in India **12**, 1–135 (2012)
17. Pakoci, E., Popović, B., Pekar, D.: Using morphological data in language modeling for serbian large vocabulary speech recognition. Computational intelligence and neuroscience **2019** (2019)
18. Pirinen, T.: Weighted finite-state methods for spell-checking and correction. Helsinki: University of Helsinki (2014)
19. Thottingal, S.: Finite state transducer based morphology analysis for Malayalam language. In: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages. pp. 1–5. European Association for Machine Translation, Dublin, Ireland (Aug 2019), https://www.aclweb.org/anthology/W19-6801