Pure and New Mathematics in AI 2025, 2(1), 10126.
https://doi.org/10.24294/pnmai10126

Article

Applications of mathematical techniques for characterization of Hindi language texts by considering the roman alphabet transforms of the texts

Hemlata Pande

Department of Mathematics, Soban Singh Jeena University, Pt. B. D. Pandey Campus Bageshwar, Uttarakhand 263642, India; hemlata.gpgc@gmail.com

CITATION

Pande H. Applications of mathematical techniques for characterization of Hindi language texts by considering the roman alphabet transforms of the texts. Pure and New Mathematics in AI. 2025; 2(1): 10126. https://doi.org/10.24294/pnmai10126

ARTICLE INFO

Received: 29 November 2024
Accepted: 25 April 2025
Available online: 12 August 2025

Abstract: The present paper is an attempt to describe the writing pattern of Hindi language texts with the help of mathematical techniques. The analyses of the selected texts have been done by the use of the roman alphabet transforms of the texts. An attempt has been made to characterize texts mathematically on the basis of the presence of different letters of alphabets and by means of quantification of the texts with the help of entropy of pattern of occurrence of letters. The characteristic curves have been formed depending on the presence of different letters in the corresponding roman text and the entropy of the pattern of occurrence of letters has also been calculated. The determined curve and the entropic extent have also been compared with the same type of curve and entropies for two texts in the English language. The work has significance in the process of language identification, as the determined curve and the specific entropic quantitative measure can be considered useful tools.

Keywords: roman alphabet transform; characteristic curves; entropy; language identification

1. Introduction

The process of analyzing "unstructured and semi-structured text data for valuable insights, trends and patterns" is known as Text Analysis [1]. "Text analysis is used in virtual assistants like Alexa, Google Home, and others." [2]. Sahu and Joshi [3] have quoted that "Natural language processing (NLP) is an approach to analyze and understand human language in a smart and useful way."

Nowadays an enormous amount of text (written in different languages) is available in digital form. In the case of any piece of written text, we can manually identify its language if we are familiar with the language. To automatically detect "the language(s) present in a document based on the content of the document "the task of 'language identification' is used [4], "Text based language identification is the task of automatically recognizing a language from a given text of document" [5]. In the research paper, [6] have worked on spoken language identification for several languages. Similarly, [7] have proposed a spoken language identification system with the help of a sequence of feature vectors.

The study about the occurrence of letters in texts and in the initial position of words of texts for the Hindi language has already been done in the previous work [8]. Similarly in the research paper of [9], the detailed study about the presence of various groups of Devanāgari symbols, according to the Phonological Inventory of Indic script for the Hindi language texts has been presented. In the present paper, an attempt has been made in the direction of identifying Hindi language texts by considering the Roman transforms of the considered texts. This work is the further extension of the earlier work [10] done with the attempt to determine a technique that can be used to compare the determined curves and the entropic extents in the case of different texts and also in the case of roman transforms of various languages.

2. Method

In the current study, following ten stories have been taken: "*Kaphan* "by 'Premchandra', "*Pinjaraa*" by 'Upendranaath Ashk', "*Taa_ii*" by 'Vishvnbharanaath Sharmaa 'Kaushik'', "*Doosaree Naak*" by 'Yashapaal', "*Raanee Ketakee Kee Kahaanee*" by 'Inshaa Allaa Khan', "*Kaisaanḍaraa Kaa Abhishaap*" by 'Agyey', "*Praayashchit*" by 'Bhagavatee Charaṇa Varmaa', "*Haar Kee Jeet*" by 'Sudarshan', "*Ek Tokaree-bhar Miṭṭee*" by 'Maadhavaraav Sapre' and "*Usane Kahaa Thaa*" by 'Chndradhar Sharmaa Guleree'. The above mentioned texts have been converted in the corresponding roman texts with the help of the tool "*Hindi to English Roman Font Converter*" available at <https://techwelkin.com/tools/hindi-to-english-roman-font-converter/>. This tool selects only a maximum of 20,000 characters for conversion. In the case of the stories that have more than 20,000 characters namely "*Kaisaanḍaraa Kaa Abhishaap*"; "*Raanee Ketakee Kee Kahaanee*"[1], the starting texts up upto the admissible limit of characters that can be converted by the tool have been taken. In the case of remaining stories (having less than 20,000 characters), the whole stories have been used. The frequency of occurrence of each letter of the English language alphabet, from *a* to *z*, has been determined for each of the above ten considered roman transforms of texts. For example, in the case of the story "*Kaphan*" the frequencies determined have been mentioned in the following table (**Table 1**).

Table 1. Frequencies of occurrence of letters *a*–*z* in the roman transform text of the story "*Kaphan*"[2].

Letter	Frequency	Letter	Frequency
a	3569	*n*	862
b	289	*o*	661
c	128	*p*	204
d	354	*q*	11
e	1663	*r*	546
f	14	*s*	411
g	238	*t*	490
h	1122	*u*	307
i	520	*v*	161
j	167	*w*	0
k	678	*x*	11
l	286	*y*	168
m	326	*z*	47

The relative frequency of each letter has been determined by dividing the frequency of the letter by the total number of alphabetical letters (*a*–*z*) occurred in the text. For example in the case of the data mentioned in the above table the relative frequency of '*a*' is 3569/13,233 = 0.2697. After determination of the relative frequencies of each of the alphabetical letters:

a) The characteristic curve for the text has been drawn as the graph of the relative frequencies of the letters of the alphabet.

b) The entropic extent of the texts regarding the presence of different letters in the roman text has been determined.

2.1. The characteristic curve

In the graph, along the horizontal axis, the alphabetical orders of letters (for ‘*a*’, order = 1, for letter ‘*b*’, order = 2, …, for letter ‘*z*’ order = 26) have been taken and along the vertical axis of the graph, the relative frequencies of corresponding letters have been mentioned. In the case of the above mentioned text “*Kaphan*” the characteristic curve for the text regarding the presence of different letters in the corresponding converted Roman text has been depicted in the following figure (**Figure 1**).

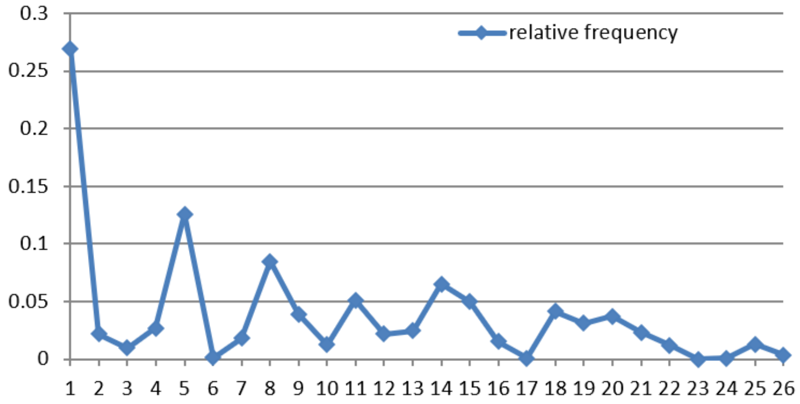

Figure 1. The characteristic curve for the story “*Kaphan*”.

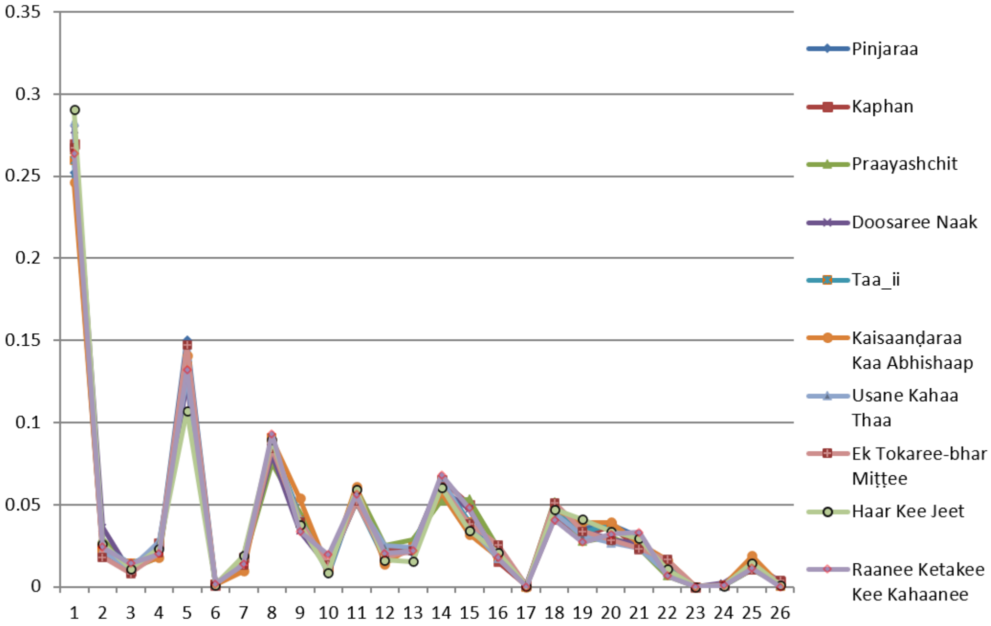

Figure 2. Characteristic curve showing the pattern of variation of relative frequencies of letters, for the roman texts of 10 Hindi language texts, when the letters are arranged in their alphabetical order.

Such characteristic curves for all considered texts/for the first 20,000 characters of texts (in case the text has more than 20,000 characters) have been obtained and have been plotted in a single graph for the purpose of the comparison of the pattern of variation of relative frequencies, where in the graph along the horizontal axis the alphabetical orders of the letters have been taken and along the vertical axis the relative frequency of occurrence of the alphabet of the corresponding order has been depicted (determined with the help of roman form). It has been seen that all the obtained characteristic curves form a similar kind of pattern (**Figure 2**); therefore, it can be said that the curve of variation of relative frequencies can be selected as the characteristic curve.

2.2. The entropic extent

"The average amount of information per symbol" is mentioned as 'entropy' in the research paper of [11]. The use of entropy in linguistic research is common (in this regard we can cite the works of [12], [13] and [14] etc.) In the research paper of [12] the entropy has been established as a measure of the average uncertainty associated with words.

In the current study, using the data for the frequencies of occurrence of various letters, the entropic measurements of the occurrence of letters in texts have been determined. The entropy (for the random variable *x*) has been calculated by the formula:

$$\text{entropy } H(x) = -\sum_{x \in X} p(x) \log_2(p(x)) \tag{1}$$

where *p*(*x*) is the probability of occurrence of the member *x* of the set of alphabet {*a*, *b*, *c*, *d*, … *z*}. For example corresponding to the data mentioned in **Table 1** the value of entropy is determined as: 3.77747.

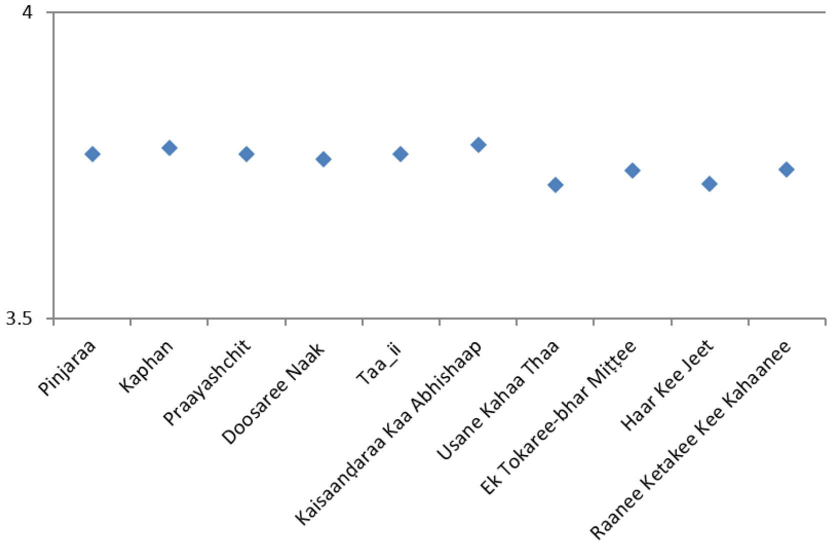

Figure 3. Entropies for the frequencies of occurrence of letters in the considered roman parts of different texts, where the entropy has been taken along the vertical axis and along the horizontal axis the texts have been depicted.

The entropy for the data of occurrence of letters in the considered roman form of various texts has been calculated and it has been seen that the values of entropy for all the considered roman part of texts are within the range from 3.7183 to 3.7832. The values of the entropy have been depicted in **Figure 3**, where entropic measurements have been marked along the vertical axis:

From **Figure 3**, it can be concluded that for all the converted roman parts of texts the entropy is in a specific range, the range can be used as the quantitative characteristic measure for the Hindi language texts.

3. Comparison of the curve for the text written in the English language

Similar kinds of characteristic curves in the case of two English texts, two short stories "Hunted Down" and "NOBODY'S STORY" written by the author: 'Charles Dickens' have been determined for comparison. The frequencies of occurrence of letters of the alphabet have been determined for the two stories and their characteristic curves have been drawn in a similar way. The pattern of characteristic curves of the above-mentioned two stories of the English language with the 10 characteristic curves determined in the case of the above-mentioned 10 texts of Hindi have been compared by their depiction in the same figure (**Figure 4**).

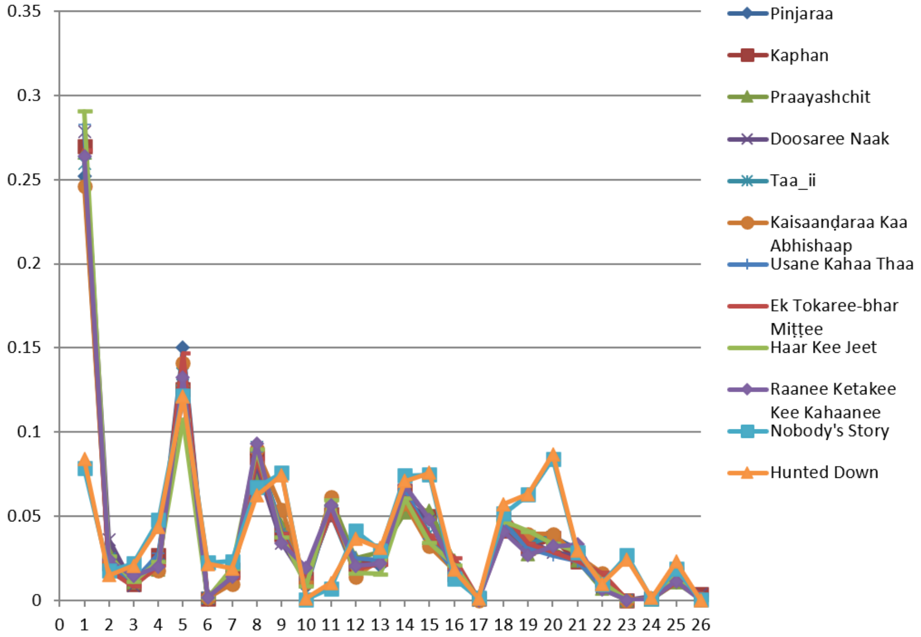

Figure 4. Comparison of the relative frequencies of occurrence of letters of the alphabet in 10 roman transformed texts of Hindi and two English short stories, where along the horizontal axis the alphabetical orders of letters and along the vertical axis the relative frequencies have been shown.

The above figure depicts that the pattern of variation of the frequencies of letters of the English language alphabet in the case of Roman transforms of Hindi language texts is not similar to the pattern for the considered English language texts while in the case of the considered roman transforms of Hindi language texts the pattern is the same for all the considered texts. Thus, it can be concluded that in the case of Hindi language texts, the characteristic curve drawn with the help of frequencies of occurrence of different letters in the corresponding roman transformed texts can be taken as a helpful tool for language characterization and language identification.

In the case of two considered English language texts the entropy corresponding to the frequency of occurrence of letters has also been calculated. The determined values were 4.1688 and 4.1781 respectively in the case of NOBODY'S STORY and Hunted Down. Thus, the entropies for the two English language texts are also not within the range as specified for the considered roman transforms of Hindi texts. Or in other words the range of entropic measure can be utilized for the language identification purpose.

4. Conclusions

In the present paper it has been concluded that for the Hindi language texts, after the conversion of texts to their corresponding roman transformed texts, the characteristic curves can be determined for the texts with the help of the frequencies of letters a-z and the determined curve can be utilized for the process of language identification; similarly the entropic measurement of frequency of occurrence of letters is also a beneficial quantitative measure for language identification purposes.

Conflict of interest: The author declares no conflict of interest.

Notes

[1] In the subsequent part of this paper, the data for the two stories namely "*Kaisaanḍaraa Kaa Abhishaap*"; "*Raanee Ketakee Kee Kahaanee*" is corresponding to their parts as considered by "Hindi to English Roman Font Converter"upto the admissible characters limit.

[2] The characters other than alphabets *a–z* (for example: *ḍ*, *ṭ* etc.) which have occurred in the roman transforms (formed with the help of considered tool) of texts have not been taken for analysis in the study.

References

1. Chen M. A Guide: Text Analysis, Text Analytics & Text Mining. Available online: https://medium.com/data-science/a-guide-text-analysis-text-analytics-text-mining-f62df7b78747 (accessed on 20 October 2024).
2. Swati. Hands-on Hindi Text Analysis using Natural Language Processing (NLP). Available online: https://www.analyticsvidhya.com/blog/2021/10/hands-on-hindi-text-analysis-using-natural-language-processing-nlp/ (accessed on 20 October 2024).
3. Sahu B, Joshi BK. A Tool for Statistical Analysis of Alphabets and Words of Hindi. In: Kumar A, Paprzycki M, Gunjan V (editors). ICDSMLA 2019. Springer Singapore; 2020.
4. Lui M, Lau JH, Baldwin T. Automatic Detection and Language Identification of Multilingual Documents. Transactions of the Association for Computational Linguistics. 2014; 2: 27-40. doi: 10.1162/tacl_a_00163
5. Indhuja K, Indu MG, Sreejith C, Raj P. Text Based Language Identification System for Indian Languages Following Devanagiri Script. International journal of engineering research and technology. 2014.
6. Singh G, Sharma S, Kumar V, et al. Spoken Language Identification Using Deep Learning. Computational Intelligence and Neuroscience. 2021; 2021(1). doi: 10.1155/2021/5123671
7. Alashban AA, Qamhan MA, Meftah AH, et al. Spoken Language Identification System Using Convolutional Recurrent Neural Network. Applied Sciences. 2022; 12(18): 9181. doi: 10.3390/app12189181
8. Pande H, Dhami HS. Mathematical Modelling of Occurrence of Letters and Word's Initials in Texts of Hindi Language. SKASE Journal of Theoretical Linguistics. 2010.
9. Pande H, Dhami HS. Analysis and Mathematical Modelling of the Pattern of Occurrence of VariousDevanāgariLetter

Symbols according to the Phonological Inventory of Indic Script in Hindi Language. Journal of Quantitative Linguistics. 2014; 22(1): 22-43. doi: 10.1080/09296174.2014.974457

10. Pande H. Applications of Mathematical Techniques for the determination of the distinctive curves for Hindi language texts. In: Pant R, Pandey V, Pandey P (editors). Artificial intelligence: a modern approach in different fields. Laxmi Book Publication; 2024.
11. Sahu B, Joshi BK. Statistical Properties of Pure Hindi and Practical Hindi. International Journal of Computer Science and Information Security. 2021. doi: 10.5281/ZENODO.5674300
12. Bentz C, Alikaniotis D. The Word Entropy of Natural Languages. Available online: https://arxiv.org/abs/1606.06996#:~:text=The%20average%20uncertainty%20associated%20with,of%20quantitative%20and%20computational%20linguistics (accessed on 20 October 2024).
13. Bentz C, Alikaniotis D, Cysouw M, et al. The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. Entropy. 2017; 19(6): 275. doi: 10.3390/e19060275
14. Arora A, Meister C, Cotterell R. Estimating the Entropy of Linguistic Distributions. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022. doi: 10.18653/v1/2022.acl-short.20