International Journal of Computational and Electronic Aspects in Engineering
Volume 3, Issue 2, June 2022, pp 13 - 17
https://doi.org/10.26706/ijceae.3.2.20220504

Phoneme Based Approach for Transliteration of Konkani Language

Sushma R. Iliger[1]
sushmariliger@gmail.com

Soniya Usgaonkar[2]
soniya@gec.ac.in

Department of Information Technology, Goa College Of Engineering, Margao, Goa, India

***Abstract* — Simultaneously with the rise of machine translation, there has been a surge in the research field of machine transliteration. Despite the fact that the two processes are distinct and serve separate purposes, transliteration aids in the optimization of machine translation models. For languages such as Arabic, Korean, Japanese, Persian, Urdu, and Hindi etc, several methodologies have been developed. In this paper we present the implementation of the phoneme-based transliteration of Konkani scripture to Roman scripture using Direct Character Mapping technique and discuss the performance with respect to the ratings from a survey conducted from a small sample of Konkani speaking individuals. From the survey conducted we obtained an average score of 3.5 with respect to word accuracy.**

Keywords—Transliteration, Phoneme based, Direct Mapping, Forward transliteration

I. Introduction

With the rise in inter-state migration in India, we now have a high number of bilingual youngsters. These youngsters learn to read, write, and communicate in the state language scriptures in which they live, but having no literary skills in their mother tongue. Despite the fact that 80 percent of migrant children proudly speak their mother tongue at home, they lack reading and writing skills in their mother tongue. This presents a communication barrier, especially in internet chatting rooms if the user is unfamiliar with the scripture, thus transliteration helps in such case.

Transliteration is the process of converting a source language scripture to a target language scripture in such a way that the transliterated output is phonologically similar to the source scripture and the target language's phonology is preserved. The distinction between transliteration and translation is that the former preserves the source scripture's phonetic qualities. It does not provide the meaning of the word, but it does provide information on how the word is pronounced in a source language, making the language accessible to those unfamiliar with the scripture but have the verbal knowledge. Whereas the latter offers the meaning of a word written in a foreign language and the phonetic qualities of the original language are not preserved. Example, for the Hindi word 'फूल' the translated English equivalent is 'Flower'. However, the transliterated English/Roman equivalent is 'Phool'.

The necessity for a transliteration system arises for a variety of reasons:

1. It is typically used to transform named entities as part of machine translation (MT) and cross-language information retrieval (CLIR).
2. It can be used to break down communication barriers in online chat rooms where someone can speak the language but is illiterate in reading or writing it.
3. It's especially useful in tourist destinations where scriptures can be read without relying on others.

Konkani is an Indo Aryan language spoken by people in Goa and parts of Maharashtra, Karnataka, and Kerala in India. As a result, the Konkani-speaking community began

Technical Article
Available online on – 02 August 2022

Cite this article – Sushma R. Iliger, Soniya Usgaonkar, "Phoneme Based Approach for Transliteration of Konkani Language", *International Journal of Computational and Electronic Aspects in Engineering*, RAME Publishers, vol. 3, issue 2, pp. 13-17, 2022.
https://doi.org/10.26706/ijceae.3.2.20220504

to use the local scriptures to write in Konkani. However, Devnagriri and Roman are the two major scriptures used in Goa, with the former being the official one. Konkani, like any other language, has a variety of dialects that are heavily influenced by the place and society in which it is spoken. In Goa, the Goan Antruz dialect has become the de facto mainstream dialect. In our research of transliteration, we have experimented with the Goan antruz dialect and results for the same have been discussed.

A. Literature Survey

Because transliteration is considered as a key component of translation, several studies have been conducted in this domain for different languages. A number of models have been created either to solve backward or forward transliteration problems. Phoneme-based transliteration models, grapheme-based transliteration models, and hybrid transliteration modes are the major techniques used to implement algorithms. Accuracy Rate is a measure used to assess the model's performance. It's the proportion of correct transliterations among the total transliterations outputted by the system.

TABLE I.
LITERATURE SURVEY OF DIFFERENT APPROACHES

Reference	***Technique***	***Approach***	***Performance (%) Metric***
[4]	Grapheme Based	Conditional Random Field based on Statistical probability	85.79 % (word accuracy)
[5]	Grapheme Based	Character Sequence Modelling (CSM), Handcrafted Rules	99.27 % (word accuracy)
[7]	Phoneme Based	Decision Tree Based	56.00 % (word accuracy)
[8]	Phoneme Based	Weighted Finite State Transducer	64 % (word accuracy)
[10]	Hybrid Based	Phonetic mapping, Rule based	96.316 % (word accuracy)

II. METHODOLOGY

A. Transliteration strategies

The building blocks of any spoken language is the unit of sound or the smallest contrastive units known as phonemes. The transliteration process in this paradigm does not use orthographic information. The source phoneme's pronunciation, rather than its letter or grapheme, is the key. Based on manually defined transcription criteria, the syllables are matched to phonemes. This model can be implemented in two ways i.e source grapheme-to-source phoneme transformation and source phoneme-to-target grapheme transformation.

The basic unit of a written language that has its own meaning or grammatical relevance is referred to as a grapheme. Transliteration done on grapheme-based techniques is considered as the process of transferring a grapheme sequence from a source language to a target language, neglecting phoneme processes. They are also known as direct approaches since there is conversion of source language graphemes into targeted language graphemes directly. To deliver any correct transliteration, these approaches expect to be well-trained using source and target transformation rules pairs, but phonetic-based methods do not. Rule based models, Statistical Machine Transliteration (SMT) based models, Finite State Transducer (FST) based models are some examples.

Hybrid transliteration techniques incorporate both grapheme and phonetic based system to make a single system. Various ways combinations can be made such as Rule based and SMT, phoneme based and SMT, phoneme based and Rule based, Rule based and HMM etc. Hybrid techniques have outperformed single systems even for languages having strong test corpora.

Since the availability of digital resources for Konkani language is low we had used the phoneme based strategy for the transliteration process.

B. Konkani Phonology

Devanagari script used for Goa Hindu Konkani has 16 vowels (V) which are called swara, 37 consonants(C). There are 2 sets of vowels, namely long and short vowels in Konkani. The long and short vowels are called vhod and san swara. In Devanagari script, a consonant without short vowel ‘a’ attached to it is called pure consonant whereas a full consonant comes with an attached ‘a’. Full consonants Devanagari form is extensively used in forward and

backward transliteration and hence we have considered it in our experiment.

It is a difficult task to analyze the pattern of distribution of vowels in any dialect. However, we made a broad observation about the frequency of vowels /a and /ā. '/a' is seen at the end of the word. '/ā' is seen in the middle of the word. Phonological rules such as reducing the high vowel before the appearance of the next vowel, omission or insertion of the final vowel, insertion or omission of a vowel in the middle of the word etc have to be followed to achieve the result with close accuracy. However, the common rules are directly taken care of when constructing a dictionary that has been used for mapping. Below is the snapshot of the devanagari vowels, matra and anusvar, consonants in Fig 2, Fig 3, Fig 1 respectively.

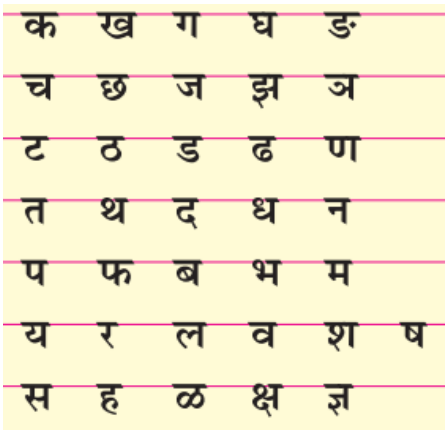

Fig. 1. Devanagari script consonants

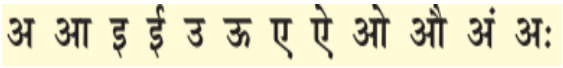

Fig. 2. Devanagari script long and short vowels

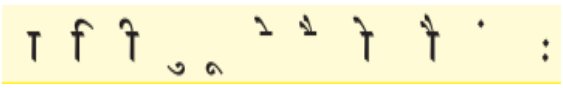

Fig. 3. Devnagiri script matras and anusvar

TABLE II

PHONETIC MAP DICTIONARY FOR VOWELS IN DEVANAGARI SCRIPTURE

Devanagari Vowels	*Devanagari Unicode*	*Roman phoneme*
अ	\u0905	a
आ	\u0906	ā
इ	\u0907	i
ई	\u0908	ī
उ	\u0909	u
ऊ	\u090a	ū
ए	\u090f	e
ऐ	\u0910	ei
ओ	\u0913	o
औ	\u0914	oh

C. Algorithm

The Overall flow can be put into the following Steps:

Step 1: Create phonetic map dictionary for Devanagari letters having Konkani phonology. The snapshot of the vowels dictionary is illustrated in Table II.

Step 2: Input Source Text (ST) I.e Konkani sentences in Devanagari Script.

Step 3: Perform tokenization on the sentences.

Step 4: Apply phonefication algorithm on each token.

Step 5: Generate equivalent target text phoneme for individual source text phoneme by direct mapping of character using the dictionary.

Step 6: Merging of transliterated phonemes to form the transliterated words. and merge transliterated words to form sentences to get Transliterated Text (TT).

Algorithm 1 Phonetic Mapping Transliteration

Input: ST
Output: TT
Initialisation :
len = ST.lenght()
for $i = 1$ to len **do**
  tokens[]=ST[i].tokenise()
end for
for x in tokens **do**
  for $i = 1$ to x.length **do**
    Generate y = getMapping(x[i])
  end for
  TT[].append(y)
end for
return TT

The following examples explain the work system:

Source Text : ' तुका मोगान सांगूंक जाय फूल दी काळजाची पाचवी ताळी चंवरतली मोन्यांनी उलयत भुलयत सर्वकाळी '

Tokenization Module : ['तुका', 'मोगान' , 'सांगूंक' , 'जाय' , 'फूल' , 'दी' , 'काळजाची' , 'पाचवी' , 'ताळी' , 'चंवरतली' , 'मोन्यांनी' , 'उलयत' , 'भुलयत' , 'सर्वकाळी']

Phonification Module :

'तुका' => त | ु | क | ा

'मोगान' => म | ो | ग | ा | न

Direct Character Module :

'तुका' => त | ु | क | ा => t | u | k | ā

'मोगान' => म | ो | ग | ा | न => m | o | g | ā |n

Merging of phonemes :

'तुका' => tukā

'मोगान' => mogān

Transliterated Text : 'tukā mogān sāngūnk jāy fūl dī kālljāchī pāchvī tāllī chnvrtlī monyānnī ulyt bhulyt srvkāllī'

III. Results

To test if the algorithm could preserve the phonetic characteristic of the source language, we used poems from the NCERT Konkani textbook of Class 1 as the data. However, the output of any transliteration process cannot be easily quantified because it depends on an individual instinct. Therefore, a survey was conducted among 50 candidates to check the accuracy of the result obtained. They were asked to rate the outputs on the scale of 1 to 5 where 1 being least accurate and 5 highly accurate. The phonetic system received an average score of 3.5 delivering satisfactory output.

Some limitations were observed in the pure phonetic system. The exact correctness of a transliteration could not be measured since it was highly dependent on a manually created dictionary. E.g.: The word 'फूल' can have transliterated English/Roman equivalents such as 'Phool', 'Phul','Fool', 'Fhul','Ful', 'Phul'. Also Perfect transliterations are in some cases, impossible when a pure phonetic based system is applied. E.g.: A name of a place, name of a person, organization, loan words were not transliterated perfectly. Country name such as 'फ्रान्स' is transliterated as 'frāns'instead of 'france'.

IV. Conclusion

Due to a lack of digital material for the source language, the majority of transliteration problems are treated with the phoneme-based technique as an initial attempt. However, it was discovered that the Phoneme based approach produces errors due to the absence of a rule-based system. Because this method relies on bilingual pronunciation data, which are not always easily available for all languages, it is extremely reliant on manually constructed dictionaries. However, this approach is appropriate for languages with little resources and the transliteration experiment has just begun.

References

[1] H. S. Priyadarshani, M. D. W. Rajapaksha, M. M. S. P. Ranasinghe, K. Sarveswaran and G. V. Dias, "Statistical Machine Learning for Transliteration: Transliterating names between Sinhala, Tamil and English," 2019 *International Conference on Asian Language Processing* (IALP), 2019, pp.244-249, https://doi.org/10.1109/ialp48816.2019.9037651

[2] Chinnakotla,Manoj K and Damani,Om P and Satoskar, "Transliteration for resource-scarce languages", *ACM*, 2010.https://doi.org/10.1145/1838751.1838753

[3] Arbabi, M.; Fischthal, S. M.; Cheng, V. C.; Bart, E.. "Algorithms for Arabic name transliteration", *IBM Journal of Research and Development*, 1994 38(2), 183–194. https://doi.org/10.1147/rd.382.0183

[4] Dhore, Manikrao & Shantanu, Kumar & Sonwalkar, Tushar. (2012). Hindi to English Machine Transliteration of Named Entities using Conditional Random Fields. *International Journal of Computer Applications.* 48. https://doi.org/10.5120/7522-0624

[5] Singh, Shailendra & Sachan, Manoj. (2019). GRT: Gurmukhi to Roman Transliteration System using Character Mapping and Handcrafted Rules. *International Journal of Innovative Technology and Exploring Engineering*. 8. 2758-2763. https://doi.org/10.35940/ijitee.i8636.078919

[6] Rajan, V.: Konkanverter - A Finite State Transducer based Statistical Machine Translitera-tion Engine Konkani Language. *Proceedings of the 5th Workshop on South and SoutheastAsian NLP, 25th International Conference on*

Computational Linguistics, pp. 11–19. Irel-and (2014). https://doi.org/10.3115/v1/w14-5502

[7] Jong-Hoon Oh and Key-Sun Choi. 2002. An English-Korean transliteration model using pronunciation and contextual rules. In *Proceedings of the 19th international conference on Computational linguistics* - Volume 1 (COLING '02). Association for Computational Linguistics, USA, 1–7. https://doi.org/10.3115/1072228.1072327

[8] Khantonthon, N., Kawtraku, A. and Poovarawan, Y. (2000), "An Enhancement of Thai Text Retrieval Efficiency by Automatic Backward Transliteration", in *proceedings of 7th International Workshop of Academic Information Networks on Systems*, Bangkok, Thailand, pp. 73-84. https://www.researchgate.net/publication/251802214_An_Enhancement_of_Thai_Text_Retrieval_Efficiency_by_Automatic_Backward_Transliteration

[9] Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL '98/EACL '98). *Association for Computational Linguistics*, USA, 128–135. https://doi.org/10.3115/976909.979634

[10] J. Nair and A. Sadasivan, "A Roman to Devanagari Back-Transliteration Algorithm based on Harvard-Kyoto Convention," 2019 *IEEE 5th International Conference for Convergence in Technology* (I2CT), 2019, pp. 1-6, https://doi.org/10.1109/i2ct45611.2019.9033576

[11] Rathod P H, Dhore M L and Dhore R M, (2013) "Hindi And Marathi To English Machine Transliteration Using SVM", *International Journal on Natural Language Computing* (IJNLC) Vol. 2, No.4, pp. 55-71. https://doi.org/10.5121/ijnlc.2013.2404

[12] TirthankarDasgupta, Manjira Sinha and Anupam Basu, "Forward Transliteration of Dzongkha Text to Braille," *Proceedings of the Second Workshop on Advances in Text Input Methods* (WTIM 2), pp. 97–106, December 2012. https://aclanthology.org/W12-4809.pdf

[13] K. Raju, T. V. Sreerekha, P. V. Vidya, R. R. Rajeev and P. C. Reghu Raj, "Tamil to Malayalam Transliteration," 2015 *Fifth International Conference on Advances in Computing and Communications*(ICACC),2015,pp.12-15,https://doi.org/10.1109/icacc.2015.86

[14] A. Murat, A. Yusup and Y. Abaydulla, "Research and Implementation of the Uyghur-Chinese Personal Name Transliteration Based on Syllabification," 2013 *International Conference on Asian Language Processing*, 2013, pp. 71-74, https://doi.org/10.1109/ialp.2013.22

[15] N. B. Jariwala and B. Patel, "Transliteration of Digital Gujarati Text Into Printable Braille," 2015 *Fifth International Conference on Communication Systems and Network Technologies*,2015,pp.572-577, https://doi.org/10.1109/csnt.2015.82