Comparison of Different Orthographies for Machine Translation of Under-Resourced Dravidian Languages

Bharathi Raja Chakravarthi
Insight Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, IDA Business Park, Lower Dangan, Galway, Ireland
https://bharathichezhiyan.github.io/bharathiraja/
bharathi.raja@insight-centre.org

Mihael Arcan
Insight Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, IDA Business Park, Lower Dangan, Galway, Ireland
michal.arcan@insight-centre.org

John P. McCrae
Insight Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, IDA Business Park, Lower Dangan, Galway, Ireland
https://john.mccr.ae/
john.mccrae@insight-centre.org

Abstract

Under-resourced languages are a significant challenge for statistical approaches to machine translation, and recently it has been shown that the usage of training data from closely-related languages can improve machine translation quality of these languages. While languages within the same language family share many properties, many under-resourced languages are written in their own native script, which makes taking advantage of these language similarities difficult. In this paper, we propose to alleviate the problem of different scripts by transcribing the native script into common representation i.e. the Latin script or the International Phonetic Alphabet (IPA). In particular, we compare the difference between coarse-grained transliteration to the Latin script and fine-grained IPA transliteration. We performed experiments on the language pairs English-Tamil, English-Telugu, and English-Kannada translation task. Our results show improvements in terms of the BLEU, METEOR and chrF scores from transliteration and we find that the transliteration into the Latin script outperforms the fine-grained IPA transcription.

2012 ACM Subject Classification Computing methodologies → Machine translation

Keywords and phrases Under-resourced languages, Machine translation, Dravidian languages, Phonetic transcription, Transliteration, International Phonetic Alphabet, IPA, Multilingual machine translation, Multilingual data

Digital Object Identifier 10.4230/OASIcs.LDK.2019.6

Funding This work was supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015 and by Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

1 Introduction

Worldwide, there are around 7,000 languages [1, 18], however, most of the machine-readable data and natural language applications are available in very few popular languages, such as Chinese, English, French, or German. For other languages resources are scarcely available and for some languages not at all. Some examples of these languages do not even have a writing system [28, 24, 2], or are not encoded in major schemes such as Unicode. The languages addressed in this work, i.e. Tamil, Telugu, and Kannada, belong to the Dravidian

languages with scarcely available machine-readable resources. We consider these languages as under-resourced in the context of machine translation (MT) for our research.

Due to the lack of parallel corpora, MT systems for under-resourced languages are less studied. In this work, we attempt to investigate the approach of Multilingual Neural Machine Translation (NMT) [16], in particular, the *multi-way* translation model [13], where multiple sources and target languages are trained simultaneously. This has been shown to improve the quality of the translation, however, in this work, we focus on languages with different scripts, which limits the application of these multi-way models. In order to overcome this, we investigate if converting them into a single script will enable the system to take advantage of the phonetic similarities between these closely-related languages.

Closely-related languages refer to languages that share similar lexical and structural properties due to sharing a common ancestor [33]. Frequently, languages in contact with other language or closely-related languages like the Dravidian, Indo-Aryan, and Slavic family share words from a common root (*cognates*), which are highly semantically and phonologically similar. Phonetic transcription is a method for writing the language in other script keeping the phonemic units intact. It is extensively used in speech processing research, text-to-speech, and speech database construction. Phonetic transcription into a single script has the advantage of collecting similar words at the phoneme level. In this paper, we study this hypothesis by transforming Dravidian scripts into the Latin script and IPA. We study the effect of different orthography on NMT and show that coarse-grained transcription to Latin script outperforms the more fine-grained IPA and native script on multilingual NMT system. Furthermore, we study the usage of sub-word tokenization [38], which has been shown to improve machine translation performance. In combination with sub-word tokenization, phonetic transcription of parallel corpus shows improvement over the native script experiments.

Our proposed methodology allows the creation of MT systems from under-resourced languages to English and in other direction. Our results, presented in Section 5, show that phonetic transcription of parallel corpora increases the MT performance in terms of the BLEU [31], METEOR [3] and chrF [32] metric [9]. Multilingual NMT with closely-related languages improve the score and we demonstrate that transliteration to Latin script outperforms the more fine-grained IPA.

2 Related work

As early as [4], researchers have looked into translation between closely-related languages such as from Czech-Russian RUSLAN and Czech-Slovak CESILKO [17] using syntactic rules and lexicons. The closeness of the related languages makes it possible to obtain a better translation by means of simpler methods. But both systems were rule-based approaches and bottlenecks included complexities associated with using a word-for-word dictionary translation approach. Nakov and Ng [30] proposed a method to use resource-rich closely-related languages to improve the statistical machine translation of under-resourced languages by merging parallel corpora and combining phrase tables. The authors developed a transliteration system trained on automatically-extracted likely cognates for Portuguese into Spanish using systematic spelling variation.

Popović et al. [34] created an MT system between closely-related languages for the Slavic language family. Language-related issues between Croatian, Serbian and Slovenian are explained by [33]. Serbian is digraphic (uses both Cyrillic and Latin Script), the other two are written using only the Latin script. For the Serbian language transliteration without

loss of information is possible from Latin to Cyrillic script because there is a one-to-one correspondence between the characters. The statistical phrase-based SMT system, Moses [23], was used for MT training in these works. In contrast, the Dravidian languages in our study do not have a one-to-one correspondence with the Latin script.

Previous proposed works on NMT, specifically on low-resource [41, 10] or zero-resource MT [20, 15], experimented on languages which have large parallel corpora. These methods used third languages as pivots and showed that translation quality is significantly improved. Although the results were promising, the success of NMT depends on the quality and scale of available parallel corpora from the pivot or third language. The third or pivot language of choice in previous works were well-resourced languages like English, German, French but many under-resourced languages have very different syntax and semantic structure to these languages. We use languages belonging to the same family which shares many linguistic features and properties to mitigate this problem. In previous works, the languages under study shared the same or similar alphabets but, in our research, we deal with the languages which have entirely different orthography.

Machine transliteration [22] is a common method for dealing with names and technical terms while translating into another language. Some languages have special phonetic alphabets for writing foreign words or loanwords. Cherry and Suzuki [11] use transliteration as a method to handle out-of-vocabulary (OOV) problems. To remove the script barrier, Bhat et al. [7] created machine transliteration models for the common orthographic representation of Hindi and Urdu text. The authors have transliterated text in both directions between Devanagari script (used to write the Hindi language) and Perso-Arabic script (used to write the Urdu language). The authors have demonstrated that a dependency parser trained on augmented resources performs better than individual resources. The authors have shown that there was a significant improvement in BLEU (Bilingual Evaluation Understudy) score and shown that the problem of data sparsity is reduced. In the work by [8], the authors translated lexicon induction for a heavily code-switched text of historically unwritten colloquial words via loanwords using expert knowledge with just language information. Their method is to take word pronunciation (IPA) from a donor language and convert them in the borrowing language. This shows improvements in BLEU score for induction of Moroccan Darija-English translation lexicon bridging via French loan words.

Recent work by Kunchukuttan et al. [27] has explored orthographic similarity for transliteration. In their work, they have used related languages which shares similar writing systems and phonetic properties such as Indo-Aryan languages. They have shown that multilingual transliteration leveraging similar orthography outperforms bilingual transliteration in different scenarios. Note that their model cannot generate translations; it can only create transliterations. In this work, we focus on multilingual translation of languages which uses different scripts. Our work studies the effect of different orthographies to common script with multilingual NMT.

3 Dravidian languages

Dravidian languages [25] are spoken in the south of India by 215 million people. To improve access to and production of information for monolingual speakers of Dravidian languages, it is necessary to have an MT system from and to English. However, Dravidian languages are under-resourced languages and thus lack the parallel corpus needed to train an NMT system. For our study, we perform experiments on Tamil (ISO 639-1: ta), Telugu (ISO 639-1: te) and Kannada (ISO 639-1: kn). The targeted languages for this work differ in several ways,

although they have nearly the same number of consonants and vowels, their orthographies differ due to historical reasons and whether they adopted the Sanskrit tradition or not [5].

The Tamil script evolved from the Brahmi script, Vatteluttu alphabet, and Chola-Pallava script. It has 12 vowels, 18 consonants, and 1 *aytam* (voiceless velar fricative). The Telugu script is also a descendant of the Southern Brahmi script and has 16 vowels, 3 vowel modifiers, and 41 consonants. The Kannada script has 14 vowels, 34 consonants, and 2 *yogavahakas* (part-vowel, part-consonant). The Kannada and Telugu scripts are most similar, and often considered as a regional variant. The Kannada script is used to write other under-resourced languages like Tulu, Konkani, and Sankethi. Since Telugu and Kannada are influenced by Sanskrit grammar, the number of characters is higher than in the Tamil language. In contrast to Tamil, Kannada, and Telugu inherits some of the affixes from Sanskrit [40, 36, 25]. Each of these has been assigned a unique block in Unicode, and thus from an MT perspective are completely distinct.

4 Experimental Settings

4.1 Data

To train an NMT system for English-Tamil, English-Telugu, and English-Kannada language pairs, we use parallel corpora from the OPUS[1] web-page [39]. OPUS includes large number of translations from the EU, open source projects, the Web, religious texts and other resources. OPUS also contains translations of technical documentation from the KDE, GNOME, and Ubuntu projects. We took the English-Tamil parallel corpora created with the help of Mechanical Turk for Wikipedia documents [35], EnTam corpus [37] and furthermore manually aligned the well-known Tamil text Tirukkural, which contains 2660 lines. Most multilingual corpora come from the parliament debates and legislation of the EU or multilingual countries, but most non-EU languages lack such resources. For our experiments, we combined all the corpus to form a **complete corpus** and split the corpora into an evaluation set containing 1,000 sentences, a validation set containing 1,000 sentences, and a training set containing the remaining sentences shown in Table 1. Following Ha et al. [16], we indicate the language by prepending two tokens to indicate the desired source and target language.

An example of a sentence in English to be translated into Tamil would be:

```
<en> <ta> Translate into Tamil
```

Table 1 Corpus statistics of the **complete corpus** (Collected from OPUS on August 2017) used for MT. (Tokens-En: Total number of tokens in the English side of parallel corpora. Tokens-Dr: Total number of tokens in the Dravidian language side of parallel corpora.)

	Number of sentences	Tokens-English	Tokens-Dravidian
English-Tamil	2,248,685	44,139,295	34,111,290
English-Telugu	224,940	1,386,861	1,714,860
English-Kannada	69,715	504,098	687,413
Total	2,543,340	46,030,254	36,513,563

[1] http://opus.nlpl.eu/

Table 2 Corpus Statistics of the **multi-parallel corpus** used for MT. (Tokens-En: Total number of tokens in the English side of parallel corpora. Tokens-Dr: Total number of tokens in the Dravidian language side of parallel corpora.)

	Number of sentences	Tokens-English	Tokens-Dravidian
English-Tamil	38,930	238,654	153,087
English-Telugu	38,930	238,654	164,335
English-Kannada	38,930	238,654	183,636
Total	116,790	715,962	501,058

Table 3 Orthographic representation of word *blue* in Tamil, Telugu and Kannada shown in native script, Latin script and IPA.

ISO 639-1	Script	Spelling	Transliteration	IPA	English
ka	Kannada	ನೀಲಿ	nili	nili	Blue
ta	Tamil	நீலம்	nilam	n̪i:lam	Blue
te	Telugu	నీలం	nilam	ni:ləm	Blue

4.2 Multi-parallel Corpus

In order to enable the training of the multi-way model, we developed a **multi-parallel corpus**, which consists of only the sentences that are available in all four languages. In this small subset of the complete corpus, most of the sentences for the Dravidian languages came from the translations of technical documents. The English sentences from the bilingual parallel corpora of three languages are aligned by collecting common English sentences from all three languages and their translation in the Dravidian languages. For the one-to-many multilingual models and many-to-one models [14], the parallel corpora were combined to form an English-to-Dravidian (Tamil, Telugu, and Kannada) NMT and Dravidian (Tamil, Telugu, and Kannada)-to-English NMT.

The corpus consists of 38,930 sentences, shown in Table 1. Combined, the corpus used to train multilingual NMT models consists of 116,790 sentences, 715,962 sources (English) tokens, and 501,058 target tokens.

4.3 Transliteration

In this section, we study the hypothesis of transliterating Dravidian scripts into the Latin script. Transliteration is a common method for dealing with technical terms and names while translating into another language. It is an approach where a word in one script is transformed into a different character set while attempting to maintain phonetic correspondence. As most of the Indian languages use different scripts, to take advantage of multilingual NMT models, we converted the Tamil, Telugu and Kannada script into the Latin script for a common representation before merging them into a multilingual corpus. We have used the Indic-trans library[2] [6] to transliterate the Dravidian side of the parallel corpus for three Dravidian languages, namely Tamil, Telugu, and Kannada, into the Latin script. The indic-trans lib produces 92.53 % accuracy for Tamil-English, 92.27 % accuracy for Telugu-English, and 91.89 % accuracy for Kannada-English.

[2] `https://github.com/libindic/indic-trans`

4.4 International Phonetic Alphabet - IPA

The International Phonetic Alphabet (IPA) [19] contains symbols for vowels, consonants and prosodic features, such stress and it is intended to be an accurate phonetic representation for all languages. We use IPA for the phonetic transcription of Dravidian languages into a single representation. We use the Epitran library [29], which is a grapheme-to-phoneme transducer supporting 61 languages. It takes the words as input and provides phonetic transcription in IPA. It has support for Tamil and Telugu but not for Kannada. Therefore, we used the Txt2ipa[3] library for Kannada, which uses a dictionary mapping to convert the Kannada script into IPA script. Table 3 shows the English word *blue* in native script, transliteration and IPA. From the figure, it is clear that the transliteration has more common sub-word units than IPA.

4.5 Translation experiments

We performed our experiments with OpenNMT [21] a toolkit for neural machine translation and neural sequence modeling. After tokenization, we fed the parallel corpora to the OpenNMT preprocessing tools i.e. OpenNMT tokenizer. Preprocessed files were then used to train the models. We used the OpenNMT parameters based on the paper [16] for training, i.e., 4 layers, 1000 for RNN size, bidirectional RNN, and 600-word embedding size, input feeding enabled, batch size of 64, 0.3 dropout probability and a dynamic learning rate decay.

The approach of [16] allows us to integrate the multilingual setting with a single encoder-decoder approach and without modification of the original OpenNMT model. This unified approach to extend the original NMT to multilingual NMT does not require any special treatment of the network during training. We compare the multilingual NMT model with bilingual models for both multilingual corpora and multiway multilingual corpora. Different evaluation sets were used for test multi-way multilingual and multilingual systems.

Table 4 Cosine similarity of the transliteration of the languages under study at character level using the **complete corpus**.

	Latin script	**IPA**
Tamil-Telugu	0.9790	0.7166
Tamil-Kannada	0.9822	0.5827
Telugu-Kannada	0.9846	0.8588

Table 5 Cosine similarity of the transliteration of the languages under study at character level using the **multi-parallel corpus**.

	Latin script	**IPA**
Tamil-Telugu	0.9867	0.6769
Tamil-Kannada	0.9825	0.5602
Telugu-Kannada	0.9855	0.5679

[3] `https://github.com/arulalant/txt2ipa`

Table 6 BLEU (B), METEOR (M) and chrF (C) scores are illustrated for systems trained with native script, Latin script and IPA. Native script is different for Tamil, Telugu and Kannada. Latin script and IPA are common script representations. Best results for each systems are shown in bold.

	Native Script			Latin Script			IPA		
	B	M	C	B	M	C	B	M	C
Bilingual systems results trained at word level									
En-Ta	**40.32**	**34.79**	**62.70**	39.7	23.48	50.10	30.67	26.37	45.27
En-Te	20.15	21.37	40.93	**20.43**	**21.42**	**41.20**	19.3	20.06	40.09
En-Kn	**28.15**	**33.53**	**60.20**	28.13	23.46	42.96	27.11	33.50	50.78
Ta-En	**32.21**	**25.65**	**44.68**	30.72	24.78	43.60	31.2	25.29	43.60
Te-En	**16.24**	**28.36**	33.22	17.96	11.84	31.26	12.65	29.23	**44.01**
Kn-En	**25.93**	**22.20**	**41.88**	23.89	20.81	39.82	20.52	18.65	17.02
Multilingual systems results trained at word level									
En-Ta	43.6	34.57	64.58	**44.23**	**35.48**	**65.02**	32.94	23.86	47.03
En-Te	23.69	23.37	42.32	**23.98**	23.93	**42.49**	22.35	**25.98**	42.86
En-Kn	28.82	33.62	**62.73**	**31.71**	35.03	46.12	30.59	**36.45**	53.94
Ta-En	29.8	24.83	46.64	**35.66**	**28.43**	**47.44**	33.86	27.34	46.89
Te-En	17.82	32.34	**56.61**	**22.95**	**24.68**	36.14	16.39	24.34	48.29
Kn-En	25.11	18.50	42.60	**28.31**	**27.63**	**42.95**	24.46	24.54	19.83

5 Results

5.1 Comparison of transliteration methodologies

While it is clear that IPA is generally a more fine-grained transliteration than the transliteration to Latin script, we wished to quantitatively evaluate this difference. Thus, we took the complete corpus for each language and for each character (Unicode codepoint) that occured in the texts, we calculated its total frequency c_f^l. We then calculated the cosine similarity between the two languages, l_1, l_2, e.g.,

$$sim^{l_1,l_2} = \frac{\sum_c f_c^{l_1} f_c^{l_2}}{\sqrt{\sum_c (f_c^{l_1})^2 \sum_c (f_c^{l_2})^2}}$$

Table 4 and 5 shows the statistics of the cosine similarity at the character level, showing that our intuition that the Latin transliteration is much more coarse-grained is well-founded as the results show that the Latin script produces a cosine similarity of about 0.98 for these three languages whereby the IPA score is lower compared to the Latin script.

To further validate this, we show in Table 3 the word *blue* in all the three languages. The root word *nil* is the same in all the languages whereby Tamil and Telugu have commonality at the whole word level. It is clear that there are far fewer commonalities in the IPA transliteration than in the Latin script transliteration.

5.2 Translation Results

Using the data, settings, and metrics described above, we investigated the impact of phonetic transcription on the machine translation of closely-related languages in multilingual NMT. We trained 54 bilingual and 18 multilingual systems corresponding to training policies and languages discussed above. All the systems were trained for 13 epochs. We use BLEU [31], METEOR [3] and chrF [32] metrics for the translation evaluation. BLEU is an

Table 7 BLEU (B), METEOR (M) and chrF (C) scores are shown for systems trained with native script, Latin script and IPA for **multi-parallel corpora with different evaluation set**. Native script is different for Tamil, Telugu and Kannada. Latin script and IPA are common script representations. Best results for each systems are shown in bold.

	Native Script			**Latin Script**			**IPA**		
	B	M	C	B	M	C	B	M	C
Bilingual systems results trained at word level									
En-Ta	31.91	22.94	43.77	**36.18**	**31.24**	**49.45**	28.67	22.92	32.35
En-Te	37.70	36.53	45.39	**38.67**	**34.12**	**48.44**	30.39	32.21	38.35
En-Kn	25.45	12.67	38.49	**26.51**	**28.66**	**39.87**	23.37	16.55	35.66
Ta-En	31.49	**37.61**	41.33	**34.75**	37.15	**43.24**	36.61	36.24	37.59
Te-En	35.30	32.23	49.35	36.44	34.69	42.72	**38.84**	**37.65**	**49.40**
Kn-En	**33.14**	21.71	44.76	30.17	**32.08**	**51.71**	24.87	18.63	45.53
Multilingual system results trained at word level									
En-Ta	37.32	38.94	50.56	**41.99**	**43.67**	49.11	38.45	39.66	**52.38**
En-Te	38.75	38.66	52.83	**39.67**	**42.75**	**56.44**	32.39	32.21	43.35
En-Kn	35.67	28.03	55.12	**37.85**	**32.43**	**60.53**	34.93	26.22	57.38
Ta-En	**36.03**	**32.32**	**54.46**	34.53	31.33	52.55	30.47	27.74	52.23
Te-En	34.22	31.17	53.14	**42.42**	**33.72**	**56.77**	30.72	25.82	52.28
Kn-En	32.15	46.65	59.49	**36.47**	**33.79**	**63.79**	34.59	41.06	56.12
Bilingual systems results trained at sub-word level tokenization									
En-Ta	36.11	20.30	53.43	**46.82**	**39.55**	**62.13**	43.63	36.36	61.90
En-Te	37.53	36.24	44.56	**39.47**	**36.34**	**58.45**	38.2	33.76	69.06
En-Kn	35.99	27.71	**55.37**	**39.20**	**42.94**	52.07	30.77	27.29	53.11
Ta-En	32.56	**23.42**	29.00	**36.62**	23.12	**44.35**	29.75	22.47	23.61
Te-En	36.12	18.93	56.63	**38.82**	**35.01**	**54.39**	39.5	25.95	37.65
Kn-En	34.85	29.26	43.86	**34.98**	**38.92**	**51.65**	33.87	24.27	45.00
Multilingual systems results trained at sub-word level tokenization									
En-Ta	39.25	31.91	**62.18**	**40.77**	**36.66**	56.52	31.34	27.32	52.16
En-Te	37.63	38.16	64.20	**38.33**	**43.34**	**67.45**	35.20	23.76	59.06
En-Kn	37.17	30.31	56.39	**37.85**	**37.08**	**59.03**	53.21	29.93	54.46
Ta-En	**37.18**	**34.69**	**57.58**	35.52	31.27	55.01	36.86	32.78	56.68
Te-En	**35.79**	**23.67**	46.76	29.61	23.28	46.97	28.43	20.39	**37.24**
Kn-En	**34.15**	39.84	62.19	30.53	**40.74**	**64.29**	27.36	24.56	29.38

automatic evaluation technique which is a geometric mean of n-gram precision. It is language-independent, fast, and shows a good correlation with human judgment. It is extensively used for various MT evaluations. The METEOR metric was designed to address the drawbacks of BLEU. We also used the chrF metric to study system output at the character level which uses F-score based on character n-grams. It is absolutely language independent and also tokenization independent.

5.2.1 Analysis of Latin script results

In order to provide a consistent evaluation of results, we wished to compare the system outputs using the native script in all settings, instead of using the output translations in IPA and Latin script. Thus, we back-transliterated the generated translations using the Indic-trans library from Latin script to native script and ran the evaluation metrics for

Table 8 Manual evaluation results of 50 sentences for translation between English and Tamil.

	Ideal	Acceptable	Possibly Acceptable	Unacceptable
Native Script				
En-Ta	8	11	14	17
Ta-En	8	13	18	11
Transliteration				
En-Ta	8	14	12	16
Ta-En	9	13	21	7
IPA				
En-Ta	6	14	17	13
Ta-En	3	18	18	11

both the corpora. Table 6 and 7 compare the results of various NMT generated translation in BLEU, METEOR, and chrF. We observe that the translations from Latin script based system provides an improvement in terms of BLEU, METEOR and chrF scores for translation from English to Tamil, Telugu, and Kannada for the bilingual systems for the multi-parallel corpus. This trend continues in the evaluation scores for the multilingual model as well. The multilingual systems outperform the baseline bilingual systems trained on the native script. The results are shown in Table 7. The METEOR and chrF score also show the same trend as the BLEU scores. Compared to the bilingual NMT system based on the native script, the multilingual NMT system based on the Latin script has improvement in the BLEU score for translation from English to Dravidian languages.

In the other direction, i.e., from Tamil, Telugu, and Kannada to English, the results are different. The Tamil→English model, based on the native script, has a higher BLEU score that the Latin Script for the multi-parallel corpus. For the Telugu→English model, based on Latin script, there is an improvement in BLEU score and Kannada-English models based on Latin script there is an improvement in BLEU score. The multilingual model of Tamil-English and Telugu-English have higher BLEU score based on the native script than the Latin script, except for the Kannada-English model where the Latin script based models outperform the native script based models. The might be the cause of translating from many languages to single languages in our case English.

5.2.2 Analysis of IPA results

To back-transcribe IPA translations into the native script, we trained an NMT system using the IPA corpus and native script corpus as a parallel resource; this was to ensure that the comparison is fair between the different transliterations. For the IPA-Tamil (Script) system, we got the 90.24 BLEU-1, and 93.07 chrF scores. BLEU-1 94.11, and chrF 94.37 for IPA-Telugu. For the IPA-Kannada BLEU-1 score was 90.51, and chrF was 89.34. We then transcribed the evaluation data to a native script using the above NMT systems. Despite the promising results in multilingual NMT, IPA results are lower compared to Latin script based systems. We observed that the scores of BLEU, METEOR, and chrF are lower than the results based on the native script in bilingual NMT translations in Table 6 and 7. It is noticeable that the scores from Dravidian languages to English trained with IPA representations did not improve the translation quality. This is due to the fact that the IPA representation was very detailed at the phonetic level than the Latin script transliteration.

5.3 Comparing BPE with word level models

There are two broad approaches to tokenize the corpora for MT. The first approach involves word level tokenization and the second is sub-word level tokenization (Byte Pair Encoding). At sub-word level, closely related languages have a high degree of similarity, thus makes it possible to effectively translate shared sub-words [26]. Byte Pair Encoding (BPE) avoids OOV issues by representing a more frequent sub-word as atomic units [38]. We train our models on space-separated tokens (words) and sub-word units. Sub-word tokenization is proven to improve the results in the translation of rare and unseen words for the language pairs like English→German, English→French and other languages [38]. Our experiments on the generated translations of the models based on the BPE corpus reveals that the systems based on Latin script have higher BLEU score in all targeted translation direction i.e. from English to Dravidian language and vice versa. Moreover, by analyzing the METEOR and chrF scores we note that systems, based on the Latin script using sub-word segmented corpora effectively reduce the translation errors. Again, we observed improvements from English into Dravidian languages but a drop in results for the other direction. Results for the model trained at the sub-word level are shown in Table 7. The transliteration-based multilingual system outperforms both the native and the IPA script based multilingual system. These results indicated that the coarse-grained transliteration to Latin script gives an improvement of MT results by better taking advantage of closely-related languages.

6 Error Analysis

We observed an improved performance of Latin script compared to native script and IPA, which is due to the limited number of characters, which better represents the phonological similarity of these languages. We see that the Latin transliteration mostly outperforms both the native script and the IPA transliteration and furthermore that the sub-word tokenization also improves performance. Surprisingly, the combination of these methodologies does not seem to be effective.

We can explain this by the example of the words ‘nilam’ and ‘nili’, which when we apply sub-word tokenization become ‘nil’ and ‘am’ or ‘i’. While Tamil and Telugu have similar morphology for this word, the common token of ‘am’ and ‘i’ are difficult to map to Kannada.

For word-level representation in native script, the number of translation units can increase with corpus size, especially for morphologically rich languages, like Dravidian languages which lead to many OOVs, and thus, a single script with sub-word units addresses the data sparsity issue most effectively.

We performed a manual analysis of the outputs generated by the different systems. Table 8 show the results of manual evaluation. We used four categories based on the work by [12]:

Ideal. Grammatically correct with all information accurately transferred.
Acceptable. Comprehensible with the accurate transfer of all important information.
Possibly Acceptable. Some information transferred accurately.
Unacceptable. Not comprehensible and/or not much information transferred accurately.

From the manual analysis, we found out that the native script and transliteration methods are more similar in terms of ideal and acceptable translation, while IPA has fewer ideal results due to errors at the character level. The unacceptable case is high in results from native script translation due to many out of vocabulary terms. All three methods have similar numbers of acceptable and possibly acceptable cases.

7 Conclusion

In this work, we described our experiments on translation across different orthographies for under-resourced languages such as Tamil, Telugu, and Kannada. We show that in the Tamil, Telugu, and Kannada to English translation direction the translation quality of bilingual NMT and multilingual NMT systems improves. In order to remove the orthographic differences between languages in the same family, we performed transcription from a native script into Latin script and IPA. We demonstrated that the phonetic transcription of parallel corpora of closely-related languages shows better results and that the multilingual NMT with phonetic transcription to Latin script performs better than IPA transliteration. This can be explained due to the coarse-grained natures of the transliteration, which produce more similarity at the character level in the target languages, which we proved by evaluating the cosine similarity of the character frequencies.

References

1 Steven Abney and Steven Bird. The Human Language Project: Building a Universal Corpus of the World's Languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 88–97. Association for Computational Linguistics, 2010. URL: http://www.aclweb.org/anthology/P10-1010.

2 Iñaki Alegria, Xabier Artola, Arantza Diaz De Ilarraza, and Kepa Sarasola. Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque, 2011.

3 Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics, 2005. URL: http://www.aclweb.org/anthology/W05-0909.

4 Alevtina Bemova, Karel Oliva, and Jarmila Panevova. Some Problems of Machine Translation Between Closely Related Languages. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*, 1988. URL: http://www.aclweb.org/anthology/C88-1010.

5 Kamadev Bhanuprasad and Mats Svenson. Errgrams - A Way to Improving ASR for Highly Inflected Dravidian Languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008. URL: http://www.aclweb.org/anthology/I08-2113.

6 Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. IIIT-H System Submission for FIRE2014 Shared Task on Transliterated Search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA, 2015. ACM. doi:10.1145/2824864.2824872.

7 Riyaz Ahmad Bhat, Irshad Ahmad Bhat, Naman Jain, and Dipti Misra Sharma. A House United: Bridging the Script and Lexical Barrier between Hindi and Urdu. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 397–408, 2016. URL: http://aclweb.org/anthology/C/C16/C16-1039.pdf.

8 Michael Bloodgood and Benjamin Strauss. Acquisition of Translation Lexicons for Historically Unwritten Languages via Bridging Loanwords. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 21–25. Association for Computational Linguistics, 2017. doi:10.18653/v1/W17-2504.

9 Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. Improving Wordnets for Under-Resourced Languages Using Machine Translation. In *Proceedings of the 9th Global WordNet Conference*. The Global WordNet Conference 2018 Committee, 2018. URL: http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_16.

10 Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. A Teacher-Student Framework for Zero-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935. Association for Computational Linguistics, 2017. doi:10.18653/v1/P17-1176.

11 Colin Cherry and Hisami Suzuki. Discriminative Substring Decoding for Transliteration. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1075. Association for Computational Linguistics, 2009. URL: http://www.aclweb.org/anthology/D09-1111.

12 Deborah Coughlin. Correlating automated and human assessments of machine translation quality. In *In Proceedings of MT Summit IX*, pages 63–70, 2003.

13 Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875. Association for Computational Linguistics, 2016. doi:10.18653/v1/N16-1101.

14 Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural, and Yoshua Bengio. Multi-way, Multilingual Neural Machine Translation. *Comput. Speech Lang.*, 45(C):236–252, September 2017. doi:10.1016/j.csl.2016.10.006.

15 Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277. Association for Computational Linguistics, 2016. doi:10.18653/v1/D16-1026.

16 Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the International Workshop on Spoken Language Translation*, 2016. URL: http://workshop2016.iwslt.org/downloads/IWSLT_2016_paper_5.pdf.

17 Jan Hajic, Jan Hric, and Kubon Vladislav. Machine Translation of Very Close Languages. In *Sixth Applied Natural Language Processing Conference*, 2000. URL: http://www.aclweb.org/anthology/A00-1002.

18 Auður Hauksdóttir. An Innovative World Language Centre : Challenges for the Use of Language Technology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA), 2014. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/795_Paper.pdf.

19 International Phonetic Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

20 Johnson, Melvin and Schuster, Mike and Le, Quoc V. and Krikun, Maxim and Wu, Yonghui and Chen, Zhifeng and Thorat, Nikhil and Viégas, Fernanda and Wattenberg, Martin and Corrado, Greg and Hughes, Macduff and Dean, Jeffrey. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. URL: http://aclweb.org/anthology/Q17-1024.

21 Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics, 2017. URL: http://www.aclweb.org/anthology/P17-4012.

22 Kevin Knight and Jonathan Graehl. Machine Transliteration. *Computational Linguistics*, 24(4), 1998. URL: http://www.aclweb.org/anthology/J98-4003.

23 Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics, 2007. URL: http://www.aclweb.org/anthology/P07-2045.

24 Steven Krauwer. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proceedings of SPECOM 2003*, pages 8–15, 2003.

25 Arun Kumar, Ryan Cotterell, Lluís Padró, and Antoni Oliver. Morphological Analysis of the Dravidian Language Family. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 217–222. Association for Computational Linguistics, 2017. URL: http://aclweb.org/anthology/E17-2035.

26 Anoop Kunchukuttan and Pushpak Bhattacharyya. Learning variable length units for SMT between related languages via Byte Pair Encoding. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 14–24. Association for Computational Linguistics, 2017. URL: http://aclweb.org/anthology/W17-4102.

27 Anoop Kunchukuttan, Mitesh Khapra, Gurneet Singh, and Pushpak Bhattacharyya. Leveraging Orthographic Similarity for Multilingual Neural Transliteration. *Transactions of the Association for Computational Linguistics*, 6:303–316, 2018. URL: http://aclweb.org/anthology/Q18-1022.

28 Mike Maxwell and Baden Hughes. Frontiers in Linguistic Annotation for Lower-Density Languages. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 29–37. Association for Computational Linguistics, 2006. URL: http://www.aclweb.org/anthology/W06-0605.

29 David R. Mortensen, Siddharth Dalmia, and Patrick Littell. Epitran: Precision G2P for Many Languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May 2018. European Language Resources Association (ELRA).

30 Preslav Nakov and Hwee Tou Ng. Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1367. Association for Computational Linguistics, 2009. URL: http://www.aclweb.org/anthology/D09-1141.

31 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002. URL: http://www.aclweb.org/anthology/P02-1040.

32 Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics, 2015. doi:10.18653/v1/W15-3049.

33 Maja Popović, Mihael Arcan, and Filip Klubička. Language Related Issues for Machine Translation between Closely Related South Slavic Languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52. The COLING 2016 Organizing Committee, 2016. URL: http://www.aclweb.org/anthology/W16-4806.

34 Maja Popović and Nikola Ljubešić. Exploring cross-language statistical machine translation for closely related South Slavic languages. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 76–84. Association for Computational Linguistics, 2014. doi:10.3115/v1/W14-4210.

35 Matt Post, Chris Callison-Burch, and Miles Osborne. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409. Association for Computational Linguistics, 2012.

36 P. Prakash and R. Malatesha Joshi. *Orthography and Reading in Kannada: A Dravidian Language*, pages 95–108. Springer Netherlands, Dordrecht, 1995. doi:10.1007/978-94-011-1162-1_7.

37 Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. Morphological Processing for English-Tamil Statistical Machine Translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122, 2012.

38 Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics, 2016. doi:10.18653/v1/P16-1162.

39 Jorg Tiedemann and Lars Nygaard. The OPUS Corpus - Parallel and Free: http://logos.uio.no/opus . In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA), 2004. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/320.pdf.

40 Devadath V V and Dipti Misra Sharma. Significance of an Accurate Sandhi-Splitter in Shallow Parsing of Dravidian Languages. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 37–42. Association for Computational Linguistics, 2016. doi:10.18653/v1/P16-3006.

41 Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575. Association for Computational Linguistics, 2016. doi:10.18653/v1/D16-1163.