Automatic Conversion of Dialectal Tamil Text to Standard Written Tamil Text using FSTs

Marimuthu K
AU-KBC Research Centre,
MIT Campus of Anna University,
Chrompet, Chennai, India.
marimuthuk@live.com

Sobha Lalitha Devi
AU-KBC Research Centre,
MIT Campus of Anna University,
Chrompet, Chennai, India.
sobha@au-kbc.org

Abstract

We present an efficient method to automatically transform spoken language text to standard written language text for various dialects of Tamil. Our work is novel in that it explicitly addresses the problem and need for processing dialectal and spoken language Tamil. Written language equivalents for dialectal and spoken language forms are obtained using Finite State Transducers (FSTs) where spoken language suffixes are replaced with appropriate written language suffixes. Agglutination and compounding in the resultant text is handled using Conditional Random Fields (CRFs) based word boundary identifier. The essential Sandhi corrections are carried out using a heuristic Sandhi Corrector which normalizes the segmented words to simpler sensible words. During experimental evaluations dialectal spoken to written transformer (DSWT) achieved an encouraging accuracy of over 85% in transformation task and also improved the translation quality of Tamil-English machine translation system by 40%. It must be noted that there is no published computational work on processing Tamil dialects. Ours is the first attempt to study various dialects of Tamil in a computational point of view. Thus, the nature of the work reported here is pioneering.

1 Introduction

With the advent of Web 2.0 applications, the focus of communication through the Internet has shifted from publisher oriented activities to user oriented activities such as blogging, social media chats, and discussions in online forums. Given the unmediated nature of these services, users conveniently share the contents in their native languages in a more natural and informal way. This has resulted in bringing together the contents of various languages. More often these contents are informal, colloquial, and dialectal in nature. The dialect is defined as a variety of a language that is distinguished from other varieties of the same language by features of phonology, grammar, and vocabulary and by its use by a group of speakers who are set off from others geographically or socially. The dialectal variation refers to changes in a language due to various influences such as geographic, social, educational, individual and group factors. The dialects vary primarily based on geographical locations. They also vary based on social class, caste, community, gender, etc. which differ phonologically, morphologically, and syntactically (Habash and Rambow, 2006). Here we study spoken and dialectal Tamil language and aim to automatically transform them to standard written language.

Tamil language has more than 70 million speakers worldwide and is spoken mainly in southern India, Sri Lanka, Singapore, and Malaysia. It has 15 known dialects [1] which vary mainly based on geographic location and religious community of the people. The dialects used in southern Tamil Nadu are different from the dialects prevalent in western and other parts of Tamil Nadu. Sri Lankan Tamil is relatively conservative and still retains the older features of Tamil[2]. So its dialect differs considerably from the dialects spoken elsewhere. Tamil dialect is also dependent on religious community. The var-

[1] http://en.wikipedia.org/wiki/Category:Tamil_dialects
[2] www.lmp.ucla.edu

Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, pages 37–45,
Baltimore, Maryland USA, June 27 2014.

iation of dialects based on caste is studied and described by A.K. Ramanujan (1968) where he observed that Tamil Brahmins speak a very distinct form of Tamil known as Brahmin Tamil (BT) which varies greatly from the dialects used in other religious communities. While performing a preliminary corpus study on Tamil dialects, we found that textual contents in personal blogs, social media sites, chat forums, and comments, comprise mostly dialectal and spoken language words similar to what one can hear and use in day-to-day communication. This practice is common because the authors intend to establish a comfortable communication and enhance intimacy with their audiences. This activity produces informal, colloquial and dialectal textual data. These dialectal and spoken language usages will not conform to the standard spellings of Literary Tamil (LT). This causes problems in many text based Natural Language Processing (NLP) systems as they generally work on the assumption that the input is in standard written language. To overcome this problem, these dialectal and spoken language forms need to be converted to Standard Written language Text (SWT) before doing any computational work with them.

Computational processing of dialectal and spoken language Tamil is challenging since the language has motley of dialects and the usage in one dialect varies from other dialects from very minimal to greater extents. It is also very likely that multiple spoken-forms of a given word within a dialect which we call as '**variants**' may correspond to single canonical written-form word and a spoken-form word may map to more than one canonical written-form. These situations exist in all Tamil dialects. In addition, it is very likely to encounter conflicts with the spoken and written-forms of one dialect with other dialects and vice versa. Most importantly, the dialects are used mainly in spoken communication and when they are written by users, they do not conform to standard spoken-form spellings and sometimes inconsistent spellings are used even for a single written-form of a word. In other words Schiffman (1988) noted that every usage of a given spoken-form can be considered as Standard Spoken Tamil (SST) unless it has wrong spellings to become nonsensical.

Few researchers have attempted to transform the dialects and spoken-forms of languages to standard written languages. Habash and Rambow (2006) developed MAGEAD, a morphological analyzer and generator for Arabic dialects where the authors made use of *root+pattern+features* representation for the transformation of Arabic dialects to Modern Standard Arabic (MSA) and performed morphological analysis. In the case of Tamil language, Umamaheswari et al. (2011) proposed a technique based on pattern mapping and spelling variation rules for transforming colloquial words to written-language words. The reported work considered only a handful of rules for the most common spoken forms. So this approach will fail when dialectal variants of words are encountered because it is more likely that the spelling variation rules of the spoken language vary from the rules of dialectal usages. This limitation hinders the possibility of the system to generalize. Alternatively, performing a simple list based mapping between spoken and written form words is also inefficient and unattainable.

Spoken language words exhibit fairly regular pattern of suffixations and inflections within a given paradigm (Schiffman, 1999). So we propose a novel method based on Finite State Transducers for effectively transforming dialectal and spoken Tamil to standard written Tamil. We make use of the regularity of suffixations and model them as FSTs. These FSTs are used to perform transformation which produces words in standard literary Tamil.

Our experimental results show that DSWT achieves high precision and recall values. In addition, it improves the translation quality of machine translation systems when unknown words occur mainly due to colloquialism. This improvement gradually increases as the unknown word rate increases due to colloquial and dialectal nature of words.

Broadly, DSWT can be used in a variety of NLP applications such as Morphological Analysis, Rule-based and Statistical Machine Translation (SMT), Information Retrieval (IR), Named-Entity Recognition (NER), and Text-To-Speech (TTS). In general, it can be used in any NLP system where there is a need to retrieve written language words from dialectal and spoken language Tamil words.

The paper is further organized as follows: In section 2, the challenges in processing Tamil dialects are explained. Section 3 explains the corpus collection and study. Section 4 explains the peculiarities seen in spoken and dialectal Tamil. Section 5 introduces the system architecture of DSWT. Section 6 describes conducted Experimental evaluations and the results. Section 7 discusses about the results and the paper concludes with a conclusion section.

2 Challenges in Processing Tamil Dialects

Tamil, a member of Dravidian language family, is highly inflectional and agglutinative in nature. The phenomenon of agglutination becomes much pronounced in dialects and spoken-form communication where much of the phonemes of suffixes get truncated and form agglutinated words which usually have two or more simpler words in them. A comprehensive study on the *Grammar of Spoken Tamil* for various syntactic categories is presented in Schiffman (1979) and Schiffman (1999). Various dialects are generally used in spoken discourse and while writing them people use inconsistent spellings for a given spoken language word. The spelling usages primarily depend on educational qualification of the authors. Sometimes, the authors intentionally use certain types of spelling to express satire and humor.

Due to this spelling and dialectal variation many-to-one mapping happens where all the variants correspond to single canonical written form. This is illustrated with the dialectal and spelling variants of the verb "paarkkiReen" (see) in Fig 1.

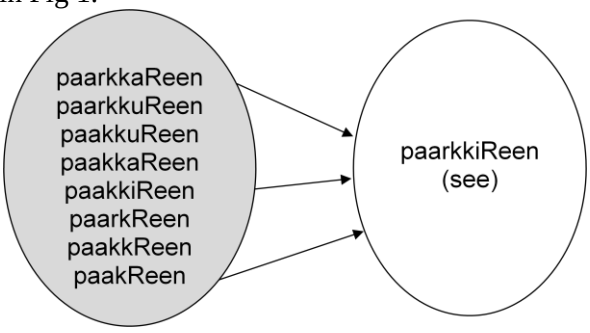

Figure 1. many-to-one mapping

For the words that belong to the above case, there is no hard rule that a particular pattern of spelling will be used and referred to while the text is written by people. In addition to this mapping, one-to-many mapping is also possible where a single spoken form maps to multiple canonical written forms.

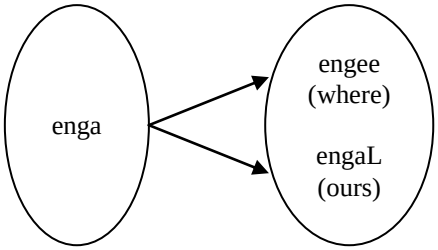

Figure 2. one-to-many mapping

In the case of one-to-many mapping, multiple written language words will be obtained. Choosing a correct written language word over other words is dependent on the context where the dialectal spoken language word occurs. In some cases, the sentence may be terminated by punctuations such as question marks which can be made use of to select an appropriate written language word. To achieve correct selection of a word, an extensive study has to be conducted and is not the focus of this paper. In the current work we are interested in obtaining as many possible mappings as possible. Many-to-one mapping occurs mainly due to dialectal and spelling variations of spoken-forms whereas one-to-many mapping happens because a single spoken-form may convey different meanings in different contexts. Dialectal spoken forms of many-to-one and one-to-many mappings are more prevalent than one-to-one mapping where a dialectal spoken form maps to exactly one written form word.

3 Data Collection and Corpus Study

The dialectal spoken form of a language is primarily used for colloquial and informal communication among native speakers. They are also commonly seen in personal blogs, social media chats and comments, discussion forums etc. Given this informal nature of the language usage, such a variety is not used in formal print and broadcasting media as they mainly use standard literary Tamil.

In our preliminary study, we found that textual contents in personal blogs, tweets, and chats have significantly large number of dialectal and spoken language words than those are found in other standard online resources such as news publishers, entertainment media websites etc.

Since we focus on processing various Tamil dialects and their spoken language variants, we have collected publicly available data from the above mentioned online resources for this work.

The collected data belongs to authors from various geographic locations where different Tamil dialects exist. The textual contents in the selected resources mainly contain movie reviews, narratives, travel experiences, fables, poems, and sometimes an informal discourse, all in a casual and colloquial manner. Further, we were able to collect variants of spoken forms which vary with respect to person, social status, location, community, gender, age, qualification etc.

Though Tamil language has 15 dialects, in this work, we focused only on 5 dialects namely, Central Tamil dialect, Madurai Tamil, Tirunelveli Tamil, Brahmin Tamil, Kongu Tamil and common spoken language forms. In Table 1, we present the corpus distribution with respect to the dialects and the number of dialectal and spoken language words.

Name of the Tamil Dialect	No. of Dialectal words
Central Tamil dialect	584
Madurai Tamil	864
Tirunelveli Tamil	2074
Brahmin Tamil	2286
Kongu Tamil	910
Common Spoken Forms	5810

Table 1. Corpus distribution among dialects

We performed an in-depth study on the collected data and found some peculiarities which exist in some dialects. Some of the observed peculiarities are described in Section 4.

4 Tamil Dialects and their Peculiarities

Some dialectal words have totally different meaning in SST and in other dialects or in standard literary Tamil. For instance, consider the following dialectal sentence (Tirunelveli Tamil)

ela, inga vaala.
Hey here come
'Hey come here!'

The words "*ela*" and "*vaala*" convey different meanings in different contexts and dialects. In SST they denote "*leaf*" and "*tail*" respectively while in Tirunelveli Tamil dialect they convey the meaning "*hey*" and "*come*" respectively.

Though these ambiguities are resolved when the context is considered, they make the transformation task challenging since this is a word-level task and no context information is taken into account during transformation.

The example in table 2, illustrates spelling based variants where the variants map to single canonical written form. We observed that the most common form of spoken-language usage is the use and representation of "*enRu*" (ADV) as four variants which are shown in Table 2.

Spoken form Variants	Written form Equivalent
[Noun/Pronoun/Verb] + "nu"	[Noun/Pronoun/Verb] + "enRu"
[Noun/Pronoun/Verb] + "nnu"	[Noun/Pronoun/Verb] + "enRu"
[Noun/Pronoun/Verb] + "unu"	[Noun/Pronoun/Verb] + "enRu"
[Noun/Pronoun/Verb] + "unnu"	[Noun/Pronoun/Verb] + "enRu"

Table 2. Spoken variants and written language

The dialectal variants of the verb "*vanthaarkaL*" (they came) is illustrated in table 3.

Dialectal variants	Written form Equivalent
[Verb] + "aaka"	[Verb] + "aarkaL"
[Verb] + "aangka"	[Verb] + "aarkaL"

Table 3. Dialectal variants & written language

It can be observed from Table 3 that the dialectal suffixes vary from each other but they all map to same written form suffix. Despite the dialectal variation, they all convey the same meaning. But they vary syntactically. The "*aaka*" suffix functions as adverbial marker in standard literary Tamil whereas it acts as person, number, gender (PNG) marker in Madurai Tamil dialect.

5 System Architecture

In this section we describe our system architecture which is depicted in Figure 3. Our dialectal spoken to written transformer (DSWT) has three main components namely, Transformation Engine, CRF word boundary identifier and heuristic Sandhi corrector.

- Transformation Engine contains FSTs for the dialectal and spoken language to standard written language transformation. The resultant words may be agglutinated and is decomposed with the help of CRF boundary identifier.
- CRF Word Boundary Identifier module identifies the word boundaries in agglutinated words and splits them into a set of constituent simpler words.
- Heuristic Sandhi Corrector module makes necessary spelling changes to the segmented constituent words and standardizes them to canonical and meaningful simpler words.

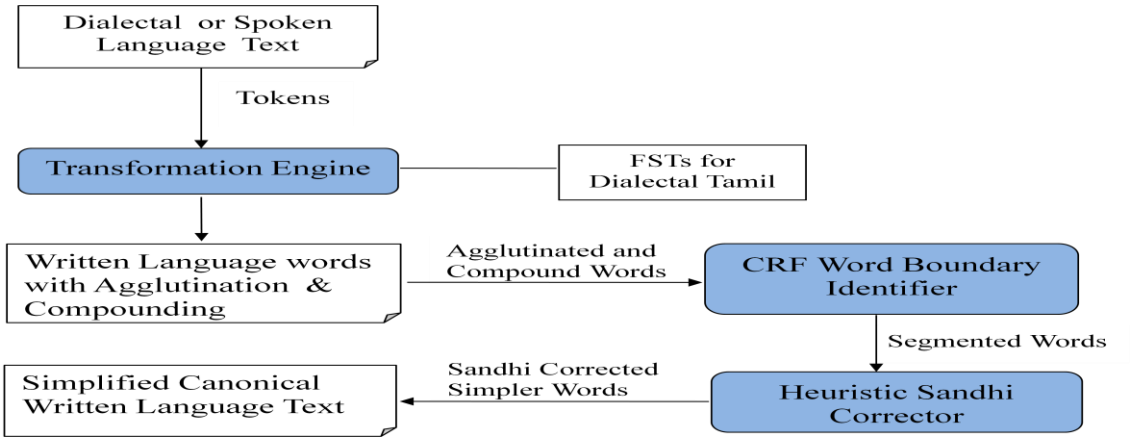

Figure 3. System Architecture

5.1 Transformation Engine

The function of Transformation engine is to transform dialectal and spoken language words into standardized literary Tamil words, similar to the official form of Tamil that is used in government publications such as official memorandums, news and print media, and formal political speeches.

Modeling FSTs for Transformation

Given the regular pattern of inflections within a paradigm, we use paradigm based approach for the variation modeling. Specifically, the dialectal usages, spoken language forms and their variants are modeled as "*root+spoken-language-suffix*" where it will get transformed into "*root+written-language-suffix*" after transformation. We had used *AT&T's FSM library*[3] for generating FSTs. The FST shown in Fig. 4 shows the state transitions for some spoken language words.

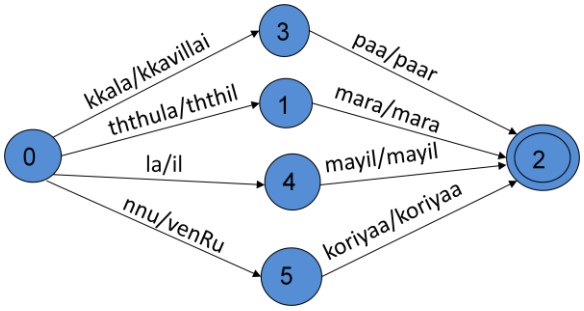

Figure 4. Sample FST

It can be observed from Figure 4 that spoken and dialectal words are processed in right to left fashion. This way of processing is adopted since the number of unique suffixation is few when compared to the number of root words. This will make the suffix matching faster and hence achieves quick transformation. This makes FSTs as an efficient tool for dialectal or variation modeling.

Algorithm for Transformation

The algorithm that is used to transform dialectal and spoken language text is given below.

1: **for** each dialectal/spoken-language word
2: check possible suffixations in FST
3: **for** each suffixation
4: *if* FST accepts & generates written language equivalents for all suffixes
5: **return (root + written-language-suffix)**
6: *else*
7: **return dialectal/spoken-language-word**
8: **for** each agglutinated & compound word
9: do CRF word boundary identification
10: **for** each constituent word (CW)
11: do Sandhi Correction
12: **return simple constituent words**

5.2 Decomposition of Agglutinated and Compound Words using CRF

Since Tamil is a morphologically rich language, the phenomenon of agglutination and compounding in standard written language Tamil is high and very common. It is also present in dialectal and spoken language Tamil. This poses a number of challenges to the development of NLP systems. To solve these challenges, we segment the agglutinated and compound words into simpler constituent words. This decomposition is achieved using two components namely

[3] http://www2.research.att.com/~fsmtools/fsm/

Agglutinated word or Compound Word	Boundary Identification and Word Segmentation	Sandhi Correction Functions			
		No Change	Insertion	Deletion	Substitution
nampuvathillaiyenRu (will not be believing)	nampuvath illai yenRu	illai	nampuvathu	enRu	
muththokuppukaLutaya (comprising of three volumes)	muth thokuppukaL utaya	thokuppukaL utaya			muu

Table 4. Boundary identification and Sandhi Correction

Table 4 clearly manifests the boundary of a constituent word within a compound or an agglutinated word which may contain one or more word-boundaries. It is observed that for "**n**" constituent words in a compound or an agglutinated word, there exists exactly (**n-1**) shared word-boundaries where (**n>0**).

CRF word boundary identifier and *Heuristic Sandhi Corrector*. We have developed the word boundary identifier for boundary identification and segmentation as described in Marimuthu et al. (2013) and heuristic rule based Sandhi corrector for making spelling changes to the segmented words.

CRF Word-Boundary Identifier

CRF based word-boundary identifier marks the boundaries of simpler constituent words in agglutinated and compound words and segments them. CRFs are a discriminative probabilistic framework for labeling and segmenting sequential data. They are undirected graphical models trained to maximize a conditional probability (Lafferty et al., 2001).

Generally word-boundary identification is studied extensively for languages such as Chinese and Japanese but the necessity for Indian languages was not considered until recently. Although there is no standard definition of word-boundary in Chinese, Peng et al. (2004) describe a robust approach for Chinese word segmentation using linear-chain CRFs where the flexibility of CRFs to support arbitrary overlapping features with long-range dependencies and multiple levels of granularity are utilized by integrating the rich domain knowledge in the form of multiple lexicons of characters and words into the framework for accurate word segmentation.

In case of Japanese, though the word boundaries are not clear, Kudo et al. (2004) used CRFs for Japanese morphological analysis where they show how CRFs can be applied to situations where word-boundary ambiguity exists.

Marimuthu et al. (2013) worked on word boundary identification and segmentation in Tamil where they model the boundary identification as a sequence labeling task [i.e. a tagging task]. The absence of word-boundary ambiguity in Tamil language favors the boundary identification task and predominantly eliminates the need for providing further knowledge to CRFs such as multiple lexicons as in the case of Chinese word segmentation. Hence we have used word level features alone for training the CRFs.

Sandhi Correction using Word-level Contextual Rules

Word-level contextual rules are the spelling rules in which each constituent word of an agglutinated or compound word is dependent either on the previous or the next or both constituent words to give a correct meaning.

After boundary identification, suppose an agglutinated or a compound word is split into three constituent words, Sandhi correction for the first constituent word is dependent only on the second constituent word while the second word's Sandhi correction depends on both first and third constituent word whereas the third constituent word's Sandhi correction depends on second constituent word alone.

Sandhi correction is performed using these rules to make necessary spelling changes to the boundary-segmented words in order to normalize them to sensible simpler words. It is accomplished using three tasks namely insertion, deletion, and substitution as described in Marimuthu et al. (2013).

For instance, after boundary identification the word *"nampuvathillaiyenRu" (will not be believing)* will be boundary marked and Sandhi corrected as shown in the Table 4 above.

Advantages of Word boundary Identification

Morphological Analysis of simpler words is much easier than analyzing agglutinated and compound words.

Tamil Dialects	No. of dialectal words	Precision (%)	Recall (%)	F-Measure (%)
Central Tamil dialect	584	88.0	89.3	88.6
Madurai Tamil	864	85.2	87.5	85.3
Tirunelveli Tamil	2074	83.4	88.6	85.9
Brahmin Tamil	2286	87.3	89.5	88.4
Kongu Tamil	910	89.1	90.4	89.7
Common Spoken Forms	5810	86.0	88.3	87.1

Table 5. Direct Evaluation Results

So the word-boundary identifier eases the task of morphological analyzer in identifying the individual morphemes. In addition, it nullifies the unknown words category if it occurs due to agglutination and compounding. As a result, it improves the recall of the morphological analyzer and any advanced NLP system. For example, with Tamil, SMT models usually perform better when the compound words are broken into their components. This 'segmentation' gives the word alignment greater resolution when matching the groupings between the two languages.

6 Experimental Evaluation

Here we perform evaluation of the performance of DSWT with test corpus of 12528 words. We perform two types of evaluations: direct and indirect evaluation.

In direct evaluation, we evaluate the system using gold standard. In indirect evaluation the system is evaluated using machine translation application. The aim in indirect evaluation is to understand the effect of dialectal and spoken language transformation in machine translation.

6.1 Direct Evaluation

We evaluate DSWT performance using the standard evaluation metrics: Precision, Recall, and F-measure. Precision and Recall values are calculated separately for each dialect using a gold standard. They are calculated using the cases described below:

A: The dialectal or spoken language transformation yields one or many correct standard written language words.
B: The dialectal or spoken language transformation yields at least one correct standard written language word.
C: The dialectal or spoken language transformation yields no output.
D: Number of dialectal or spoken language words given as input.

Precision is then calculated as: A/**(**D-C)
Recall is calculated as: (A+B)/D
F-Measure is the harmonic mean of *Precision* and *Recall.*
The obtained results for the considered 5 Tamil dialects and common spoken language forms are summarized in Table 5 above.

6.2 Indirect Evaluation

For indirect evaluation, we had used DSWT with Google Translate (GT) to measure the influence of DSWT in Tamil-English machine translation, and evaluated the improvement.

Our test data had 100 Tamil sentences which are of dialectal and colloquial in nature. At first, we used GT to translate these sentences to English. This is Output1. Then we used our DSWT to transform the dialectal sentences into standard written Tamil. After this, the standard sentences were translated to English using GT. This corresponds to Output2.

We then performed subjective evaluations of Output1 and Output2 with the help of three native Tamil speakers whose second language is English. The three evaluation scores for each sentence in Output1 and Output2 are averaged. The obtained scores are shown in Table 6.

Subjective Evaluation Scores before dialectal Transformation		**Subjective Evaluation Scores after dialectal Transformation**	
No. of sentences	*Achieved Scores*	*No. of Sentences*	*Achieved Scores*
20	0	4	0
70	1	14	1
8	2	28	2
2	3	30	3
0	4	24	4

Table 6. Subjective evaluation results

We used a scoring scale of 0-4 where
0 → no translation happened.

Before performing Dialectal Transformation Task		**After performing Dialectal Transformation Task**	
Dialectal Spoken Tamil	*Google Translate results*	*Standardized Written Tamil*	*Google Translate results*
ஒடனே வந்துரு. (otanee vanthuru)	vanturu otane. (✘)	உடனே வந்துவிடு. (utanee vanthuvitu)	Come immediately. (✓)
ஒடனே வந்துருல. (otanee vanthurula)	vanturula otane. (✘)	உடனே வந்துவிடு. (utanee vanthuvitu)	Come immediately. (✓)
அவங்க வந்தாங்க. (avanga vanthaanga)	she had come. (**?**)	அவர்கள் வந்தார்கள். (avarkaL vanthaarkaL)	They came. (✓)
அவுக வந்தாக. (avuka vanthaaka)	avuka to come. (✘)	அவர்கள் வந்தார்கள். (avarkaL vanthaarkaL)	They came. (✓)

Table 7. Tamil-English Google Translate results before and after dialectal text transformation

Sentences marked as (✘) are incorrectly translated into English and those that are marked as (**?**) may be partially correct. The sentences that are marked as (✔) are the correct English translations.

1 → lexical translation of few words happen and no meaning can be inferred from the translation output.
2 → complete lexical translations happen and some meaning can be inferred from the translation output.
3 → meaning can be inferred from translation output but contains some grammatical errors.
4 → complete meaning is understandable with very minor errors.

It can be observed from the results in Table 6 that GT failed to translate dialectal and spoken language sentences. But the failure got mitigated after transformation causing dramatic improvement in translation quality. The following Table illustrates few examples where the translation quality has improved after transforming dialectal spoken language.

It must be noted from Table 7 that after the transformation of dialectal spoken language, all the sentences were able to achieve their English equivalents during machine translation. This suggests that almost all word categories in Tamil can achieve improved translations if the words are given as standard simple written language words. This experiment emphasizes the importance of feeding the machine translation systems with standard written language text to achieve quality translations and better results.

7 Results and Discussion

We observe that the achieved accuracy is higher for Kongu Tamil dialect when compared to other dialects. This is because words in this dialect are rarely polysemous in nature. But the number of polysemous words is high in the case of Madurai and Tirunelveli Tamil dialect and this resulted in low accuracy of transformation.

While performing transformation, the possible causes for ending up with unknown words may be due to the absence of suffix patterns in FSTs, errors in input words, uncommonly transliterated words, and English acronyms. The standard written language words convey a particular meaning in standard literary Tamil and completely different meaning in dialectal usages. For instance, consider the verb "*vanthaaka*". In standard literary Tamil, this is used in the imperative sense "*should come*" while in Tirunelveli Tamil dialect it is used in the sense "*somebody came*".

8 Conclusion and Future Work

We have presented a dialectal and spoken language to standard written language transformer for Tamil language and evaluated its performance directly using standard evaluation metrics and indirectly using Google Translate for Tamil to English machine translation. The achieved results are encouraging.

There is no readily available corpus for processing dialectal and spoken Tamil texts and we have collected the dialectal and spoken language corpus for developmental and evaluation tasks. This corpus can be made use of for developing other NLP applications.

In case of one-to-many mapping, multiple written language forms will be emitted as outputs. Hence, determining which written-form of word to be adopted over other resultant written-forms has to be done based on the meaning of the whole sentence in which the spoken-language word occurs. This will be the focus of our future direction of the work.

References

A. K. Ramanujan. 1968. *Spoken and Written Tamil, the verb*. University of Chicago. Pages 74.

Fuchun Peng, Fangfang Feng and Andrew McCallum. 2004.*Chinese Segmentation and New Word Detection using Conditional Random Fields*, Computer Science Department Faculty Publication Series. Paper 92. University of Massachusetts – Amherst.

Harold F. Schiffman. 1979. *A Grammar of Spoken Tamil*, Christian Literature Society, Madras, India. Pp. i-viii, 1-104.

Harold F. Schiffman. 1988. *Standardization or restandardization: The case for "Standard" Spoken Tamil*, Language in Society, Cambridge University Press, United States of America. Pages 359-385.

Harold F. Schiffman. 1999. *A Reference Grammar of Spoken Tamil*, Cambridge University Press, Pp. i-xxii,1-232.

J. Lafferty, A. McCallum, and F. Pereira. 2001. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* In Proceedings of the 18th International Conference on Machine Learning, pages 282–289.

Marimuthu K., Amudha K., Bakiyavathi T. and Sobha Lalitha Devi. 2013. *Word Boundary Identifier as a Catalyzer and Performance Booster for Tamil Morphological Analyzer*, in proceedings of 6th Language and Technology Conference, Human Language Technologies as a challenge for Computer Science and Linguistics, Poznan, Poland.

Milton Singer and Bernard S. Cohn. 2007. *The Structure of Variation: A Study in Caste Dialects*, Structure and Change in Indian Society, University of Chicago, Chapter 19, pages 461-470.

Nizar Habash and Owen Rambow. 2006. *MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects*, In proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, Australia. Pages 681-688

Sajib Dasgupta and Vincent Ng. 2007. *Unsupervised Word Segmentation for Bangla*, In proceedings of the Fifth International Conference on Natural Language Processing(ICON), Hyderabad, India.

Taku Kudo, Kaoru Yamamoto and Yuji Matsumoto. 2004. *Applying Conditional Random Fields to Japanese Morphological Analysis*, In proceedings of Empirical Methods on Natural Language Processing, Barcelona, Spain.

Umamaheswari E, Karthika Ranganathan, Geetha TV, Ranjani Parthasarathi, and Madhan Karky. 2011. *Enhancement of Morphological Analyzer with compound, numeral and colloquial word handler*, Proceedings of ICON-2011: 9th International Conference on Natural Language Processing, Macmillan Publishers, India.