Parameswari Krishnamurthy*

Development of Telugu-Tamil Transfer-Based Machine Translation System: An Improvization Using Divergence Index

<https://doi.org/10.1515/jisys-2018-0214>

Received January 22, 2018; previously published online November 6, 2018.

Abstract: Building an automatic, high-quality, robust machine translation (MT) system is a fascinating yet an arduous task, as one of the major difficulties lies in cross-linguistic differences or divergences between languages at various levels. The existence of translation divergence precludes straightforward mapping in the MT system. An increase in the number of divergences also increases the complexity, especially in linguistically motivated transfer-based MT systems. This paper discusses the development of Telugu-Tamil transfer-based MT and how a divergence index (DI) is built to quantify the number of parametric variations between languages in order to improve the success rate of MT. The DI facilitates MT in proposing where to put efforts for the given language pair to attain better and faster results. In addition, handling strategies of different types of divergences in a transfer-based approach to MT are discussed. The paper also includes the evaluation method and how an improvization takes place with the application of DI in MT.

Keywords: Machine translation, transfer-based machine translation, divergence, divergence index, Telugu, Tamil.

2010 Mathematics Subject Classification: 68T50.

1 Introduction

Machine translation (MT) is one of the challenging tasks in natural language processing, as it is a highly knowledge-intensive activity. It requires different kinds of knowledge, such as linguistic, pragmatic, world, cultural, and domain knowledge, to understand and process a text from one language to another. In MT, there are a number of methods that are being practiced all over the world; chiefly, they are direct, interlingual, transfer-based methods, and a combination of these besides the statistical, corpus-based methods and the recent developments such as neural MT and recurrent neural networks. This paper discusses the improvization of transfer-based Telugu-Tamil MT system, which was developed under the Indian language to Indian language (IL-IL) MT project¹ using the divergence index (DI). It attempts to experimentally explain how to build a DI for Telugu to Tamil MT by listing all possible linguistic differences that may occur at any level (surface, shallow, intermediate, deep, and deeper levels) of linguistic analysis, and how DI is helpful to improvise different MT tools in order to obtain the expected result. In addition, handling strategies of different types of divergences in the transfer-based approach to MT are discussed. The paper also includes the evaluation method and how an improvization takes place with the application of DI in MT.

¹ IL-IL MT is a consortium project funded by the Department of Electronics and Information Technology, Ministry of Communications and Information Technology, Government of India, under the project name Sampark.

*Corresponding author: Parameswari Krishnamurthy, Assistant Professor, Centre for Applied Linguistics and Translation Studies, University of Hyderabad, Hyderabad-500 046, India, e-mail: parameshkrishnaa@gmail.com

2 Telugu-Tamil MT

Telugu and Tamil are major Dravidian languages with rich literary tradition, sharing indubitable linguistic similarities and dissimilarities. Tamil belongs to the South Dravidian, whereas Telugu belongs to the South Central Dravidian branch of the Dravidian family [10]. Though enormous efforts are exerted in building MTs for Indian languages, most of the MT systems are developed for English to Indian languages, and vice versa. The English to Tamil MT activities include Anuvadakh (English to Indian language MT system) and the works of Soman and Menon [24], Poornima et al. [18], Saravanan [22], Pandian and Kadhivelu [16], Ramasamy et al. [20], Kumar et al. [12], and Rajeswari et al. [19], to name a few. Similarly, English-Telugu MT systems have been developed by the Centre for Applied Linguistics and Translation Studies – University of Hyderabad, Anitha and Kommaluri [1], and Keerthi et al. [9], to mention but a few. However, not many efforts are put forth to develop an IL-IL MT.

Considering the fact that Indian languages share a number of features [5, 14, 25], high-quality MT between them would be definitely achievable. The Telugu-Tamil transfer-based MT system is an assembly of various linguistic modules run on specific engines whose output is sequentially maneuvered and modified by a series of modules until the output is generated. It employs a three-stage architecture: stage 1, source language (SL) analysis; stage 2, SL to target language (TL) transfer; and stage 3, TL generation.

The most crucial linguistic modules in SL analysis include a morphological analyzer, parts of speech (POS) tagger, chunker, named entity recognizer, and simple parser. The SL to TL transfer module includes the multi-word expression (MWE) component, transfer grammar (TG) component, and lexical transfer (LT) component consisting of synset and bilingual lexicons. TL generation includes agreement (AGR) modules (consisting of inter-chunk and intra-chunk AGR modules) and a morphological generator (MG). All these modules have been integrated on the platform called *Dashboard* based on blackboard architecture [17], which configures data flow in a specified pipeline, as given in Figure 1.

Figure 1: Telugu-Tamil MT Architecture.

3 Translation Divergence

The term “translation divergence” refers to distinctions or differences that occur between languages when they are translated, and affects the “well-formedness” of the TL. Translation divergence occurs when the underlying concept or “gist” of a sentence is distributed over different words or in different configurations for different languages [4]. The notion of divergence in MT is comparable to the linguistically motivated notion of parametric variation, i.e. cross-linguistic distinctions.

3.1 Dorr’s Divergence

Dorr classified cross-linguistic distinctions into several categories of divergences across English, Spanish, and German to account for divergences in the UNITRAN (UNiversal TRANslator) MT system. She divided divergences into two major types: syntactic and lexical-semantic divergences. Syntactic divergences are majorly divided into five: constituent order, preposition stranding, long-distance movement, null subject, and dative divergences. Lexical-semantic divergences are divided into seven: conflation, structural, thematic, categorical, demotional, promotional, and lexical divergences. Dorr’s description of divergences is restricted to English, Spanish, and German. There can be many other divergences when different language pairs are involved, and the study of divergence needs further exploration. Following Dorr, a few divergence studies have explored Indian languages, such as English-Sanskrit-Hindi MT [6], English-Sanskrit MT [15], English-Hindi MT [3, 7], Sanskrit-Hindi MT [23], Hindi-Telugu MT [21], and English-Bengali MT [2], to name a few. In most of the cases, Dorr’s divergences are noticed as rare phenomena and do not pose similar problems as far as Telugu and Tamil are considered. Further, Dorr discussed the divergences involved in interlanguage-based MT, i.e. UNITRAN; however, transfer-based MT requires further exploration.

3.2 Telugu-Tamil Divergence

The current research attempts to classify divergences into three major kinds: morphological, syntactic, and lexical-semantic divergences.

A. Morphological Divergence

Morphological divergence refers to differences that are found in inflectional and derivational (only productive) devices of words between a pair of languages in MT. Open word class categories such as nouns, verbs, and adjectives and closed word classes such as pronouns, number words, and nouns of space and time (NST) are studied between Telugu and Tamil to find out morphological divergences. Uninflected word classes, i.e. indeclinables and non-productive derivational word forms, are excluded here because they are listed in the lexicon and a straightforward mapping between them solves the problem in MT. For example, nouns in Telugu and Tamil are major word classes inflecting for number and case. The divergences in these classes are exemplified below.

Example A.1: Number Marking

There are two numbers in Telugu and Tamil, viz. singular and plural. Singular has no particular distinguishing marker, and plural is marked with the basic allomorph *-ka* in Tamil and *-lu* in Telugu. In Tamil, the plural marking is found obligatory with rational nouns² and optional with irrational nouns³ [13]. In Telugu, the plural marking is usually obligatory with nouns to denote a plurality (see Table 1).

The divergence in number inflection is expressed as

$$\text{Te.}^4 \text{ NN}[-\text{rational}] - \langle \text{plural} \rangle \Rightarrow \text{Ta.}^5 \text{ NN}[-\text{rational}] - (\langle \text{plural} \rangle).$$

² Rational nouns denote masculine and feminine, which are regarded as human.

³ Irrational nouns denote animals and inanimate things, which are regarded as neuter.

⁴ Telugu.

⁵ Tamil.

Table 1: Number Marking in Nouns.

Number	Telugu	Tamil	Gloss
Singular (rational)	<i>ammāyi</i>	<i>peṇ</i>	“girl”
Plural (rational)	<i>ammāyi-lu</i>	<i>peṇ-kaḷ</i>	“girls”
Singular (irrational)	<i>illu</i>	<i>viṭu</i>	“house”
Plural (irrational)	<i>il-lu</i>	<i>viṭu(-kaḷ)</i>	“houses”

Example A.2: Case Markers and Postpositions

Telugu and Tamil use a wide variety of case markers and postpositions to indicate various syntactico-semantic relations between nouns and verbs. The major inflectional differences in case marking occur due to two reasons: (i) the choice of items in terms of inflections, viz. the oblique stem formation, case, and postposition, and (ii) the order of their presentation. For instance, an example for the direction to an interior location is shown in Table 2, where the order of suffixes in Telugu and Tamil may differ.

This divergence is explicated as below:

Te. Noun- ±Number- ±Stem-formative- ±Postposition- ±Case.
 Ta. Noun- ±Number- ±Stem-formative- ±Case- ±Postposition.

B. Syntactic Divergence

Syntactic divergence here refers to syntactic structural differences that occur between pairs of languages. Divergences due to case mismatches, agreement, anaphora, negation, subordination, and clitics are noticed. Various syntactic processing and a robust TG are obviously required to overcome syntactic divergence. Some examples of syntactic divergences between Telugu-Tamil are given below.

Example B.1: Case Marker

Each case marker has a number of functions, and it is obvious that they lead to case mismatches in MT. For instance, Telugu and Tamil agree in the usage of dative case marker in various functions, viz. the beneficiary of an action, the goal of motion, and the experiencer subject [10, 26], among other functions. However, to express a possessive relationship between two inanimate nouns, one of the nouns of inanimate category carries the dative marker to express the locative function in Telugu. Here, the dative case marker relates two noun phrases that have the holonymic (HOL)-meronymic (MER) relationship. The word that is a holonym takes the dative case marker in Telugu and, on the contrary, the locative case marker is in use in Tamil as in the following example.

- (1) Te. *gōḍa- ku kiṭikī uM- di.*
 wall- DAT window.NOM be.PRS- 3.SG.N.
 Ta. *cuvarr- il Jannal iru- kkir- atu.*
 wall- LOC window.NOM be- PRS- 3.SG.N.
 ‘The wall has a window.’

Divergence rule:

Table 2: Postpositions in Telugu and Tamil.

No.	PSP	Telugu	Tamil	Gloss
1.	Locative:interior:direction	<i>iMṭi- lōpali- ki/</i> <i>iMṭi- lō- ki</i> house.OBL- inside- DAT	<i>viṭṭ- ukk- uḷ(ḷē)/</i> house.OBL- DAT- inside <i>viṭṭ- iṇ- uḷ(ḷē)</i> house.OBL- GEN- inside	“inside the house”

Te. NN[+HOL]-DAT NN[+MER] => Ta. NN[+HOL]-LOC NN[+MER].

Example B.2: Infinitive Construction

The use of infinitive construction in Tamil is wide ranging, whereas its use is restricted and obsolete in Telugu except in the augmentation of the main verb with an auxiliary verb [11, 13]. Infinitives (INF) are used as complements to different clauses in Tamil, whereas those in Telugu may use gerundivals (GER).

For instance, the infinitive clause occurs as a complement to desiderative verbs such as *virumpu* “want,” *ācaippau* “desire,” etc. in Tamil. In this respect, the dative-marked gerundivals act as a complement in Telugu as explicated in the example below.

(2) Ta. *nāṇ [cinimā. v- ukku.p pō.k- a] virumpu- kīr- ēṇ.*
I cinema- DAT go- INF want- PRS- 1.SG.

Te. *nēnu [cinimā- ku vell- aḍāni- ki] iṣṭapaḍu- tunnā- nu.*
I cinema- DAT go- GER- DAT want- PRS- 1.SG.
'I want to go to the cinema.'

Divergence rule:

Te. VM-GER-DAT VM[+desiderative] => Ta. VM-INF VM[+desiderative].

C. Lexical-Semantic Divergence

Lexical-semantic translation divergences are characterized by properties that are entirely lexically determined between languages. A concept expressed by a lexeme may not have the similar meaning in all contexts. Lexical-semantic divergence in Telugu and Tamil occur mainly due to lexical ambiguities and MWEs.

Example C.1: Lexical Ambiguities

For example, a lexeme that is used to express a concept in a language may not have the same meaning in all contexts. When it has multiple meanings, word sense disambiguation is required to overcome lexical ambiguity and to select an appropriate sense with its form in the TL.

For instance, the lexeme *kuṭṭu* in Telugu is ambiguous and expresses three different senses as given below.

Sense 1: *kuṭṭu* “to bite,” as in the context of *cīma* “an ant,” etc. The equivalent word in Tamil is *kaṭi* “to bite.”

Sense 2: *kuṭṭu* “to stitch,” as in the context of *baṭṭalu* “clothings.” The equivalent word in Tamil is *tai* “to stitch.”

Sense 3: *kuṭṭu* “to pierce,” as in the context of *cevulu* “ears” or body parts, etc. The equivalent word in Tamil is *kuttu* “to pierce.”

Example C.2: MWEs An MWE is a collocation of words that often come together with non-compositional semantics. These forms are sequences of two or more words that generally express a co-occurrence meaning. Telugu and Tamil show a number of conflation divergences as shown in Table 3.

Table 3: Example for Conflational Divergence in Telugu and Tamil.

No.	Telugu	Tamil	Gloss
1.	<i>kobbari nīru</i>	<i>iḷanīr</i>	“coconut water”
2.	<i>janāBā</i>	<i>makkaḷ tokai</i>	“population”
3.	<i>elluMḍi</i>	<i>nāḷai marunāḷ</i>	“day after tomorrow”
4.	<i>ākali avvu</i>	<i>paci</i>	“to feel hungry”
5.	<i>snānaM ceyyi</i>	<i>kuḷi</i>	“to take bath”
6.	<i>pariśōDiMcu</i>	<i>āyvu cey</i>	“to research”

4 Divergence Index

DI represents a measure of the differences that occur between languages. The variations in linguistic features can be seen in any level (L) (surface, shallow, intermediate, deep, and deeper levels). These levels are identified as L1 (surface level), L2 (shallow level), L3 (intermediate level), L4 (deep level), and L5 (deeper level), according to depth of variation. Identifying the divergence with its level between a pair of languages enables computing and quantifying the effort that is required to build an MT. DI uses a list of linguistic features to identify and classify divergences exhaustively into different levels in order to understand its depth.

Table 4 provides instances where divergences are possible with reference to a given feature in the said languages. Y indicates the presence of a linguistic feature in a language, and N indicates its absence. When the two languages share similar features [see Table 4 (1.) and (2.)], it means no divergence (indicated by 0). When they differ [see Table 4 (3.) and (4.)], there arises divergence (indicated by 1).

In certain cases, Y/N is given to indicate optional use of a feature. When both the SL and the TL show optional, it means no divergence [see Table 4 (5.)]. When only the SL shows optional, it is counted as divergence because the TL element may not be directly mapped when the option differs [see Table 4 (6.) and (7.)]. When the option occurs only in the TL, it is counted as no divergence [see Table 4 (8.) and (9.)] because the TL optionally behaves like the SL; hence, the SL features can be directly mapped to the TL.

4.1 Morphological DI

In Table 4, morphological DI is marked. For example, the number inflection of nouns (see Table 1) can be explained with different levels.

L1 (surface level) checks whether nouns in both languages inflect for number. If so, L2 (shallow level) indicates whether the number distinctions, such as singular and plural, are realized or not. L3 (intermediate level) inspects for the distinct marker for singular and plural. L4 (deep level) identifies differences in the number marking, if any, as explained in Table 1. That is, the distinction between rational and irrational nouns may play a role in their inflection of plural marking. L5 (deeper level) checks for the deeper linguistic feature in which the languages show any other difference.

Table 5 and Figure 2 illustrate the morphological divergence quantification for different linguistic levels.

Table 4: DI Table.

No.	SL feature	TL feature	DI
1.	Y	Y	0
2.	N	N	0
3.	Y	N	1
4.	N	Y	1
5.	Y/N	Y/N	0
6.	Y/N	Y	1
7.	Y/N	N	1
8.	Y	Y/N	0
9.	N	Y/N	0

Table 5: Morphological DI Quantification.

Category	L1	L2	L3	L4	L5
Nouns	0	5	22	124	0
Verbs	0	12	33	51	0
Adjectives	0	9	0	0	0
Pronouns	0	12	0	0	0
Number words	0	1	7	8	0
NST	0	1	6	0	0
Total	0	40	68	183	0

Figure 2: Morphological Divergence.

Table 6: Syntactic DI Quantification.

Category	L1	L2	L3	L4	L5
Case markers	16	34	27	3	3
Postpositions	0	9	13	0	0
Agreement	9	14	9	10	8
Anaphora	0	2	8	13	8
Negation	0	0	8	6	1
Subordination	0	11	24	8	0
Clitics	4	5	0	0	0
Total	29	75	89	40	20

Figure 3: Syntactic Divergence.

4.2 Syntactic DI

The syntactic divergence is quantified for different levels as expressed in Table 6 and Figure 3.

4.3 Lexical-Semantic DI

The approximate lexical-semantic divergence is calculated based on the entries in synset lexicons available in both languages. Ambiguous words are studied, and about 6706 words show ambiguity in Telugu when they are mapped with Tamil. About 6798 MWEs are identified in Telugu and are given appropriate equivalents in the TL, Tamil.

5 Handling Strategies

Based on the results of DI, the morphological and syntactic divergences between Telugu and Tamil are approximately 43.47% and 43.24%, respectively. Various handling strategies are proposed in this section to overcome divergences in the current MT system. The modules that are majorly involved in handling divergences are the parser, TG, LT, MWE component, AGR modules, and MG.

5.1 Handling Morphological Divergence

Morphological divergences are handled by the following modules.

(i) **Parser:** Telugu-Tamil MT employs a simple parser expressing two types of relations. It shows *kāraka* (K) relations of nouns with respect to verbs and other *non-kāraka* (R) relations between other constituents. *kāraka* is the one that performs an action and has various participants to execute the action. An action in a sentence is denoted by a verb and participants are through nouns, and the relationship between nouns and a verb is a K relation. Apart from K relations, a sentence contains other types of relations as well. These relations are marked by other relations (R), such as constituents expressing purpose, reason, genitive, etc. For instance, Telugu uses a similar case marker, i.e. *-tō*, to denote instrumental and associative cases. Parser finds out the exact case name based on the context using the database of rules. It makes use of ontological features of words to disambiguate their roles in sentences.

(ii) **TG:** TG in the current MT is equipped to solve certain morphological divergences. It contains a list of rules matching SL morphological features and transfers them into the expected TL. For instance, in Telugu, the dative case marking with NST denotes the direction when they occur with verbs of motion. In Tamil, the dative case marker is disallowed in such construction.

A TG rule to handle this divergence is written as below.

```
R13: NP((%NST<lcat="nst", case="o", cm="ki">))
      VGF((%VM<root="$motion", lcat="v">)) =>
      NP((%NST<lcat="nst", case="d", cm="0">))
      VGF((%VM<root="$motion", lcat="v">)).
```

TG rules contain SL features on the left side and TL features on the right side with the delimiter => in the middle. It gets information from the morphological analyzer, POS tagger, chunker, and parser outputs of SL, and transfers them into the TL.

(iii) **LT:** LT solves certain morphological divergences. Other than lexical words, LT also contains a list of functional words and their equivalents in the TL. For instance, the divergence due to auxiliaries in compound verb formation is handled by LT by substituting the SL auxiliary to the expected TL auxiliary.

(iv) **AGR modules:** Morphological divergence due to the agreement is solved by AGR modules. For instance, the gender distribution in the singular is different in Telugu and Tamil. AGR modules aim at providing the exact TL gender.

(v) **MG:** MG plays an important role in handling most of the morphological divergences in the current MT architecture. Once the SL morphological features are transferred to the TL by LT, the MG generates well-formed word forms of TL.

(vi) **Postprocessor:** Postprocessor is also involved in providing provision for solving certain morphological divergence. For instance, the number word “one” has special adjectival forms such as *oru* (used before consonants) and *ōr* (used before vowels) in Tamil. However, in Telugu, it does not have this distinction. In this case, postprocessor selects an appropriate number word “one” by looking at the initial sound of the following words.

5.2 Handling Syntactic Divergence

Syntactic divergences are mainly handled by TG. The input for TG is the shallow parsed SL text. As mentioned, TG rules contain two components: the SL structure on the left side and the expected TL structure on the right side. The following major tasks are handled by TG [8].

(a) **Insertion:** Insertion of a new node is possible by TG indicated by the symbol “+.” For instance, Tamil optionally uses copula verb in verbless constructions. To insert a copula, the following rule is executed:

```
R1: NP(({%NN}{%SYM<root="&dot;", lcat="punc">})) =>
      NP(({%NN}{}) + VGF<root="āku", lcat="v">
      ((+{%VM<root="āku", lcat="v">}+
      %SYM<root="&dot;", lcat="punc">}))
```


As seen in R1, a new node VGF with the copula verb *āku* is inserted in TL. The agreement features of copula are provided by AGR modules.

(b) **Deletion:** Deletion in TG rule empties nodes. For instance, in R1, the symbol SYM in NP is deleted and inserted again in VGF. Once a node is deleted, it is shown by {}.

(c) **Modification:** Modification in TG involves a change in the node features of TL. For example, when the predicate expresses capability mood, the subject is marked with the instrumental case marker (*-āḷ*) in Tamil, whereas the subject in Telugu is in the nominative case (O). This is written as a TG rule below:

```
R2: NP<lcat="n",case="d",cm="O", drel="k1:$x">({%NN})*)
VGF<lcat="v",tam="a_gala_$a">({%VM}) =>
NP<lcat="n",case="o",cm="āḷ", drel="k1:$x">({%NN})*)
VGF<lcat="v",tam="a_gala_$a">({%VM}).
```

(d) **Reordering:** Reordering of nodes in TG is possible. For example, in Telugu, a reduplicated noun would have a dative marked on the second segment for distributive sense, whereas in Tamil, the dative is marked on the first segment of the noun.

Te. *iMt(i) iMtī- kī* “each house”

Ta. *vīṭṭ- ukku vīṭu* “each house”

```
R3: NP(({%NN<root="$x", lcat="n", case="o", cm="O">}
{%NN<root="$x", lcat="n", case="o", cm="ki">}))
=> NP(({%NN<root="$x", lcat="n", case="o", cm="ki">}
{%NN<root="$x", lcat="n", case="o", cm="O">})).
```

(e) **File handling:** A list of items as files is maintained to write a global rule. For instance, nouns [+body part] occur with the postposition *paina/ mī.da* to denote surface location in Telugu. Such nouns are marked with the locative case marker in Tamil. A file with the list of body parts is maintained to handle such divergences.

```
V1: R4:: "$x=bodyparts.txt"
R4: NP(({%NN<root="$x", cm="paina">})) =>
NP(({%NN<root="$x", cm="il">})).
```

5.3 Handling Lexical-Semantic Divergence

Lexical-semantic divergences are mainly handled by two modules, viz. MWE component and LT. A word with a number of senses will definitely make a problem in MT. In Telugu and Tamil, it is common to find out ambiguous words, and they need to be systematically resolved to obtain the appropriate equivalent. An exhaustive set of TG rules operating on the identification of the ambiguous words and disambiguating them by looking at the subject or the object nouns as suggested above is built. For instance, the following TG rules are samples to handle the different senses of Telugu word *kuṭṭu* in Tamil:

```
V1: R1:: "$x=animate.txt"
R1: NP<root="$x", lcat="n"> VGF<root="kuṭṭu", lcat="v"> =>
NP<root="$x", lcat="n"> VGF<root="kaṭi", lcat="v">
```

```
V2: R2:: "$y=inanimate.txt"
R2: NP<root="$y", lcat="n"> VGF<root="kuṭṭu", lcat="v"> =>
NP<root="$y", lcat="n"> VGF<root="tai", lcat="v">
```

```
V3: R3:: "$z=bodyparts.txt"
R3: NP<root="$z", lcat="n"> VGF<root="kuṭṭu", lcat="v"> =>
NP<root="$z", lcat="n"> VGF<root="kuttu", lcat="v"> .
```

6 Evaluation

The human evaluation method, i.e. the native speaker's judgment, is adopted to evaluate the Telugu-Tamil MT output. Each output sentence is given a score based on the scale given in Table 7.

Parameters such as fluency and comprehensibility are calculated based on the score given to the MT output. Fluency is defined as the percentage of the total number of sentences with a score between 3 and 4 out of the total number of sentences with the highest score (i.e. 4). Comprehensibility is defined as the percentage of the total number of sentences with a score between 2 and 4 out of the total number of sentences with the highest score.

Let the total number of sentences be S .

Let the total number of sentences with highest score be $S * 4 = N$.

$$\text{Fluency} = \sum_{i=3}^4 Si/N. \quad (1)$$

$$\text{Comprehensibility} = \sum_{i=2}^4 Si/N. \quad (2)$$

Sentences with a score of 1 and 0 are considered failures in the output. Hence, they do not contribute to calculating the best performance of the system.

Information about input sentences and evaluators are given below:

Number of sentences taken for evaluation = 500.

Number of tokens = 5234.

Domains = tourism, health, and general story (sentences are taken from the Internet, such as Wikipedia and blogs written in Telugu).

Number of native speakers involved in the evaluation = 3.

Profession of evaluators = professor (age = 64 years), home maker (age = 48 years), and student (age = 21 years).

The average results are presented before divergent rules (Table 8) and after divergent rules (Table 9).

Table 7: MT Human Evaluation Scale.

Sentence output quality	Score
Type-A: perfect translation	4
Type-B: clear and understandable (with minor error)	3
Type-C: understandable (with minor error)	2
Type-D: not understandable or has major error	1
Type-E: nonsense	0

Table 8: Telugu-Tamil MT Evaluation-1.

Parameters	Tourism	Health	General story
Fluency	64.46%	62.2%	63.0%
Comprehensibility	65.34%	63.76%	65.2%

Table 9: Telugu-Tamil MT Evaluation-2.

Parameters	Tourism	Health	General story
Fluency	88.66%	84%	87.45%
Comprehensibility	90.66%	87.33%	90%

- (i) **Before divergence rules:** Using the DI database, the TG is enhanced with 442 rules and again the system is evaluated for its fluency and comprehensibility.
- (ii) **After divergence rules:** As shown in Table 9, the system performance is rather improved with the application of rules derived from the DI discussed.

7 Conclusion and Future Work

In this paper, building of a DI is introduced to improvise the transfer-based Telugu-Tamil MT, which helps in improving the success rate of MT. The building of DI requires a detailed linguistic study and can be applied to various areas of linguistic research other than MT. For instance, in the area of language teaching, DI may be useful as it gives a contrastive analysis of two languages. As proved in this paper, DI may be helpful for improvization of any linguistic research where two languages are involved. A more thorough study is required to test the feasibility and usage of DI in different areas with different languages.

Bibliography

- [1] N. Anitha and V. Kommaluri, SMT using Joshua: an approach to build 'enTel' system, *Lang. India Spec. Vol. Probl. Parsing Indian Lang.* **11** (2011), 1–6.
- [2] N. S. Dash, Linguistic divergences in English to Bengali translation, *Int. J. Engl. Linguist.* **3** (2013), 31–40.
- [3] S. Dave, J. Parikh and P. Bhattacharyya, Interlingua-based English-Hindi machine translation and language divergence, *Mach. Transl.* **16** (2001), 251–304.
- [4] B. J. Dorr, *Machine translation: a view from the Lexicon*, MIT Press, Cambridge, MA, 1993.
- [5] M. B. Emeneau, India as a linguistic area, *Language* **32** (1956), 3–16.
- [6] P. Goyal and R. M. K. Sinha, Translation divergence in English-Sanskrit-Hindi language pairs, in: *Sanskrit Computational Linguistics*, vol. 5406, pp. 134–143, Springer, Berlin, Heidelberg, 2009.
- [7] D. Gupta and N. Chatterjee, Identification of divergence for English to Hindi EBMT, in: *Proceeding of MT Summit-IX*, pp. 141–148, 2003.
- [8] IL-ILMT Consortium, Indian language to Indian language machine translation system: software requirement specifications, in: *ILMT Consortium, Hyderabad*, <http://researchweb.iiit.ac.in/rashid.ahmedpg08/ilmt.html>, 2008.
- [9] L. Keerthi, E. R. Lakshmi and L. R. Theja, Rule-based machine translation from English to Telugu with emphasis on prepositions, in: *First International Conference on Networks & Soft Computing (ICNSC2014)*, 2014.
- [10] B. Krishnamurti, *The Dravidian languages*, Cambridge University Press, Cambridge, 2003.
- [11] B. Krishnamurti and J. P. L. Gwynn, *A grammar of modern Telugu*, Oxford University Press, Delhi, 1985.
- [12] A. M. Kumar, V. Dhanalakshmi, K. P. Soman and S. Rajendran, Factored statistical machine translation system for English to Tamil language, *Pertanika J. Soc. Sci. Hum.* **22** (2014), 1045–1061.
- [13] T. Lehmann, *A grammar of modern Tamil*, Pondicherry Institute of Linguistics and Culture, Pondicherry, 1993.
- [14] C. P. Masica, *Defining a linguistic area: South Asia*, University of Chicago Press, Chicago, IL, 1976.
- [15] V. Mishra and R. B. Mishra, Study of example based English to Sanskrit machine translation, *Polibits* **37** (2008), 43–54.
- [16] L. S. Pandian and K. Kadhivelu, Machine translation from English to Tamil using hybrid technique, *Int. J. Comput. Appl.* **46** (2012), 36–42.
- [17] K. Pawan, A. K. Rathaur, A. Rashid, K. S. Mukul and S. Rajeev, Dashboard: an integration & testing platform based on black board architecture for NLP applications, in: *Proceedings of 6th International Conference on Natural language Processing and Knowledge Engineering (NLP-KE)*, Beijing, China, August, 2010.
- [18] C. Poornima, V. Dhanalakshmi, M. Anand Kumar and K. P. Soman, Rule based sentence simplification for English to Tamil machine translation system, *Int. J. Comput. Appl.* **25** (2011), 38–42.
- [19] S. Rajeswari, P. Sethuraman and K. Krishnakumar, English to Tamil machine translation system using universal networking language, *Sādhana* **41** (2016), 607–620.
- [20] L. Ramasamy, O. Bojar and Z. Žabokrtský, Morphological processing for English-Tamil statistical machine translation, in: *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pp. 113–122, 2012.
- [21] G. U. M. Rao, M. Christopher, B. B. Madhavi and S. K. Pandey, Transfer of agreement in Hindi-Telugu in machine translation system in: *9th International Conference on South Asian Languages (ICOSAL)*, Punjab University, Punjab, 2010.
- [22] S. Saravanan, *English to Tamil machine translation: rule based approach*, LAP LAMBERT Academic Publishing, Germany, 2012.

- [23] P. Shukla, D. Shukl and A. Kulkarni, Vibhakti divergence between Sanskrit and Hindi, in: *Proceedings of the International Sanskrit Computational Linguistics Symposium*, pp. 198–208, Springer, 2010.
- [24] K. P. Soman and A. G. Menon, English to Tamil machine translation system, in: *9th Tamil Internet Conference (INFITT)*, Chemmozhi Maanaadu, Coimbatore, India, 2010.
- [25] K. V. Subbarao, *South Asian languages: a syntactic typology*, Cambridge University Press, Cambridge, 2012.
- [26] M. K. Verma and K. P. Mohanan, eds., *Experiencer subjects in South Asian languages*, Center for the Study of Language (CSLI), Stanford, 1990.