A TAMIL - TELUGU MACHINE TRANSLATION SYSTEM

Uma Maheshwar Rao, G.[1], Parameshwari, K.[2], Sreenivasulu, N.V.[3], Christopher Mala[4] and Krupanandam, N.[5]

Centre for Applied Linguistics and Translation Studies
UNIVERSITY OF HYDERABAD

Abstract: We present the development of Machine Translation (MT) System which translates texts from Tamil to Telugu and vice-versa (Bi-directional). It is based on Transfer Approach. The System's Architecture is divided into three stages i.e. Source language Analysis module (SL), Source language to Target language Transfer module (SL-TL) and Target language generation module (TL). The major cross linguistic differences that are found between Tamil and Telugu while building Machine Translation system are discussed here.

1. Introduction: The development of Machine Translation (MT) is one of the most challenging tasks of Natural Language Processing applications. Even though Tamil and Telugu are two closely related languages belonging to Dravidian language family, still they exhibit a considerable amount of diversity at every level viz. morphological, syntactic, semantic and lexical levels. Keeping these in mind, building a Machine Translation System for this language pair can be not only non-trivial but also challenging. The paper deals with the issues in the development of an automatic Tamil-Telugu Machine Translation[1] System which is being developed under the project of IL-IL MT at CALTS. To build a more sophisticated and effective Machine Translation system, it is significant to identify divergences (i.e. cross linguistic differences) between the pair of languages . The Divergences between Tamil-Telugu are discussed below.

2. Divergences between Telugu-Tamil: Translation divergence occurs when the underlying concept or 'gist' of a sentence is distributed over different words or different configurations for different languages (Dorr,1993). In Machine Translation, identifying such variation is crucial to obtain qualitatively the right output.The two major levels of divergences may be noticed as, syntactic divergences and lexico-semantic divergences.

3. Syntactic Divergences: In Telugu-Tamil, major syntactic divergences occur in the following cases.

3.1. Case syncretism and mismatches: Case syncretism occurs when a single inflected form corresponds to two or more case functions (Comrie 1991: 44-47). In Telugu and Tamil, a number of instances of case syncretisms are found.

1. Dative Swapping: In Tamil, a noun when inflected for dative may take a locative postposition which in the target language Telugu will have the swapped structures as in the following:

(1) Ta. *kumār vīṭṭu-**kku** uḷḷē ce-ṉṟ-āṉ.* Eng. Kumar went inside the house.
'Kumar house-**dat** inside go-PST-3sm'
Te. *kumār iMṭi lō-**ki**/lōpali-**ki** veḷḷ-āḍu.*
'Kumar house inside-**dat**/inside-**dat** go-PST-3sm'

2. Dative as possession/location: The nouns of inanimate category when inflected for locative in Tamil express the *part-whole relationship* (cf. Subbarao & Bhaskararao, 2004). On the contrary, Telugu uses a dative case marker (see (2)).

(2) Ta. *cuvaṟṟ-**il** jaṉṉal iru-kkiṟ-atu.* Eng. The wall has a window.
'wall-**loc** window be-PRS_3sn'
Te. *gōḍa-**ku** kiṭikī uM-di.*
'wall-**dat** window be-PRS_3sn'

3. Dative as Genitive: The nominal complement of the postposition is

[1]. This Machine Translation system is developed as a part of the Consortium of Indian Languages to Indian languages Machine Translation Systems funded by DIT, Ministry of Information Technology, Government of India.

inflected for dative to the use of it as a genitive in the adnominal usage in Tamil whereas the corresponding dative is optional in Telugu (see (3)).

(3) Ta. *pūmi-**kku**/pumi.y-iṉ aṭi.y-il nīr uḷḷatu.* Eng. The water is at the bottom of the earth.
'earth-**dat**/earth-gen bottom-loc water be-PRS-3sn'
Te. *Bhūmi-**(ki)** aḍugu-na nīru uMdi.*
'earth-**(dat)** bottom-loc water be-PRS-3sn'

4. Dative as the expression of specific time: The Tamil noun marked for dative case expresses the specific time limit or duration of time roughly equivalent to the English 'within' whereas Telugu does not. (see (4))

(4) Ta. *eṉ-akku pattu nāṭ-kaḷ-**ukk**uḷ puttakatt-ai.k koṭuṅkaḷ.* Eng. Give me the book within ten days.
'me-dat ten day-pl-**dat**.inside book(obl)-Acc give-imp[hon]'
Te. *naa-ku padi rōju-la lōpu pustākaM ivvaMḍi.*
'me-dat ten day-pl inside book give-imp[hon]'

5. Double Dative Construction: A dative predicate may assign a locative case involving part relationship in Tamil, whereas Telugu uses dative. (see (5))

(5) Ta. *avaṉ /avaṉ-ukku kaṇṇ-**il** aṭipaṭ-ṭ-atu.* Eng. 'He got hurt in his eyes.'
'he-obl/he-dat eye-**loc** get-hurt-pst-3.sg.n.'
Te. *vāḍi /vāḍi-ki kanṭi-**ki** debbai tagiliMdi.*
'he-obl/he-dat eye-**dat** injury touch-pst-3.sg.n. (Subbarao & Bhaskararao 2004:169)

6. Accusative in Dative Subject Construction: In Dative subject construction, the direct object is inflected for accusative when the predicate is realized as verbs of cognition in Tamil, whereas Telugu object remains in nominative case. (see (6))

(6) Ta. *sītā.v-ukku eṉ.ṉ-**ai**.t teri.y-um.* Eng. Sita knows me.
'Sita-dat I-**acc** know-3.sg.n'
Te. *sīta-ki nēnu-**ø** telusu.*
'Sita-dat I -**nom** know'

7. Case Assigned By Postpositions: There are postpositions in both the languages, which assign case to their complements (noun phrase) which are morphologically manifested. (see (7))

(7) Ta. *kumār uṉṉ-**ai** nōkki.p pō-ṉ-āṉ.* Eng. Kumar went towards you.
'Kumar you-acc towards go-pst-3p.sg.m'
Te. *kumār nī kēsi veḷḷ-ā-ḍu.*
'Kumar you-**obl** towards-dat go-pst-3p.sg.m'

The other examples are,

Telugu	Tamil	Meaning
N_NOM + tappa	N_ACC + tavira	'except'
N_OBL+ lāMṭi	N_ACC + pōl/pōla	'like'
N_OBL + guriMci	N_ACC+ kuṟittu	'about'
N_ACC + baṭṭi	N_GEN + pati	'accordingly' and etc.,

3.2. Lexical Passive Construction: While expressing the potential mood of capability or ability, the subject in Tamil is marked by the instrumental case (*āl)* or in lexical passive (Subbarao & Bhaskararao, 2004) and Telugu in the nominative case (Ø).

(8) Ta. *eṉṉ-**āl** teluku naṉṟāka pēc-a-muṭi.y-um.* Eng. I can speak Telugu well.
'I - **by** Telugu well speak-inf-can-3s.sg.n'
Te. *nēnu-**ø** telugu bāgā māṭlāḍ-a-gala-nu.*
'I-**nom** Telugu well speak-inf-can-1p.sg'

3.3. Complementizer: In Telugu, the complementizer *ani* is dropped optionally when it accompanies the verb *anu* 'say'. But it is mandatory in Tamil for

embedded clause construction with the corresponding verb.

(9) Ta. *nāṉ uṉṉ-ai yārō* ***eṉṟu*** *niṉai-tt-ēṉ.* Eng. I thought of you as someone else.
'I you-Acc who-dub **COMP** think-pst-1.sg.'
Te. *nēnu nin-nu evarō* ***(ani)*** *anuk-unn-ā-nu.*
'I you-Acc who-dub **(COMP)** think-pst-1.sg.'

Tamil permits the infinitive form of *eṉ* 'tell' in the same interpretation of *eṉṟu*. But this is disallowed in Telugu.

(10) Ta. *nāṉ avaṉ nallavaṉ* ***eṉṟu/eṉ-a.c*** *co.ṉ-ṉ-ēṉ.* Eng. 'I told that he is a good man.'
'I he oodman **COMP/ tell-inf** tell-pst-1.sg.'
Te. *nēnu vādu maMcivādu* ***ani/*an-a*** *ceppā-nu.*
'I he goodman **COMP/ *tell-inf** tell-pst-1.sg.

However, *eṉa* in Tamil does not inflect for any clitics, whereas *eṉṟu* can inflect for.

(11) Ta. *nāṉ avaṉ nallavaṉ* ***eṉṟ-e/*eṉa.v-e*** *co.ṉ-ṉ-ēṉ.* Eng. I told him only as a good man'.
'I he goodman **COMP-emp/*tell-inf-emp** tell-pst-1.sg.'
Te. *nēnu vādu maMcivādu* ***an-e*** *ceppā-nu.*
'I he goodman **COMP-emp** tell-pst-1.sg.'

3.4. Verbal Reflexive: In Tamil, the use of verbal reflexive (VR) is optional in simple clause constructions (cf. Lehmann, 1989: 361), whereas it is mandatory in Telugu.

(12) Ta. *avaṉ taṉṉ-aik kaṇṇāṭi.y-il pār-ttu.k-****koṇ****-ṭ-āṉ/pār-tt-āṉ.* Eng.He saw himself in the mirror.
'he self-acc mirror-loc see-conj.par-**VR**-pst-3.sg.m/ see-pst-3.sg.m'
Te. *vāḍu tana-ni addaM-lō cūsu-****konn****-ā-ḍu/*cus-ā-ḍu.*
'he self-acc mirror-loc see-**VR**-pst-3.sg.m/* see-pst-3.sg.m'

The use of VR is optional in Tamil, when the nominal reflexive occurs. But the sentence will be ambiguous when both nominal and verbal reflexives are omitted. For instance,

(13) Ta. *avaṉ kaṇṇāṭi.y-il pār-tt-āṉ .*
'he mirror-loc see-pst-3.sg.m'
Eng. He saw others/himself in the mirror.

However, in certain constructions Tamil disallows verbal reflexives as follows,

(14) Ta. *avaṉ kuḻantai.y-ai muttamiṭṭ-ā-ṉ.* Eng. He kissed the child.
'he child-ACC kiss-pst-3.sg.m'
Te. *vāḍu pillavādi-ni muddupeṭṭu-****kon****.n-ā-ḍu.* (p.c., Subbarao)
'he child-ACC kiss-**VR**-pst-3.sg.m'

There are certain verbs in Tamil and Telugu which are inherently reflexive. For instance, *oppukkoḷ* 'admit', *nāṉṟukoḷ* 'hang (oneself)', *paṅkukoḷ* 'participate' etc., in Tamil and *paḍukonu* 'sleep', *oppukonu* 'admit', *mēlkonu* 'be awake', etc. in Telugu.

3.5. Gerund vs. Infinitive: The use of the infinitive construction in Telugu is obsolete (Cf. Krishnamurti, 1985; Uma Maheshwar Rao, 2002) except in the augmentation of main verb with auxiliary verb. In the desiderative clause, Tamil uses an infinitive whereas Telugu constructs the sentence with a gerund.

(15) Ta. *nāṉ uṉakku coll-****a*** *maṟa-nt-ēṉ.* Eng. I forgot to tell you.
'I you-dat tell-**inf** forget-pst-1.sg.'
Te. *nēnu nī-ku cepp-****aḍaM*** *maricipō.y-ā-nu.*
'I you-dat tell-**ger** forget-pst-1.sg.'

In purposive clauses, Telugu uses gerund inflected with Dative whereas Tamil uses an infinitive.

(16) Ta. *nāṉ paṭikk-****a*** *va-nt-ēṉ.* Eng. I came to study.
'I read-**inf** come-pst-1.sg.'
Te. *nēnu cadav-****aḍāni-ki*** *vacc-ā-nu.*
'I read-**ger-dat** come-pst-1.sg.'

To express the negation in the present tense, Telugu uses the gerund.

(17) Ta. *nāṉ ippōtu va-r**a**.v-illai.* Eng. I do not come now.
'I now come-**inf**-not'
Te. *nēnu ippuḍu rāv-**aḍaM** lēdu.*
'I now come-**ger** not'

3.6. Agreement

1. Quantifier-Noun Agreement: Inanimate nouns need not necessarily agree with the quantifiers in Tamil, whereas in Telugu they do.

(18) Ta. *pattu$_i$ rūpāy iru-kkiṟ-**atu$_j$**.* Eng. I have ten rupees.
'ten repee-sg be-prs-**3.sg.n**'
Te. *padi$_i$ rūpāya-lu unn-ā-yi$_i$.*
'ten repee-pl be-prs-**3.pl.n**'

The uncountable (mass) nouns like *pālu* 'milk', *biyyam* 'rice', *nīḷḷu* 'water' etc., are inherently plural in Telugu and verb displays agreement whereas in Tamil, it is in singular.

(19) Ta. *kumār vīṭṭ-il niṟaiya pāl$_i$ iru-kkiṟ-**atu$_i$**.* Eng. Kumar have plenty of milk in home.
'Kumar home-loc plenty milk be-pre-**3.sg.n**'
Te. *kumār iMṭ-lō cālā pālu$_i$ unn-ā-**yi$_i$**.*
'Kumar home-loc plenty milk be-pre-**3.pl.n**'

2. Subject-Nominal Predicate Agreement: Telugu nouns when occur as nominal predicates, agree in number and person with their subjects whereas in Tamil, there is no such agreement.

(20) Ta. *nāṉ$_i$ oru peṇ-**ø$_j$**.* Eng. I am a girl.
'I a girl'
Te. *nēnu$_i$ oka ammāyi-**ni$_i$**.*
'I a girl-**1.sg**'

Similarly, Telugu predicate adjectives agree in number and person.

(21) Ta. *nāṉ$_i$ nalla.v-**aṉ$_j$**.* Eng. I am a good man.
'I good-**3.sg.m**'
Te. *nēnu$_i$ maMci-vāḍi-**ni$_i$**.*
'I good-3.sg.m-**1.sg**'

3. Subject-Verbal Predicate Agreement: Tamil has a three way gender distinction in singular as masculine, feminine and neuter whereas Telugu shows a two way gender distinction in singular as masculine vs. non-masculine.

(22) Ta. *avaḷ$_i$ va-nt-**āḷ$_i$**.* Eng. She came.
'she come-pst-**3.sg.f**'
Te. *āme$_i$ vacc-iM-**di$_i$**.*
'she come-pst-**3.sg.nm**'

(23) Ta. *nāy$_i$ va-nt-**atu$_i$**.* Eng. The dog came.
'dog come-pst-**3.sg.n**'
Te. *kukka$_i$ vacc-iM-**di$_i$**.*
'dog come-pst-**3.sg.nm**'

3.2. Lexico-semantic divergences: Lexico-semantic translation divergences are accounted for by means of parameterization of the lexicon. (Dorr, 1993:20)

3.2.1. Conflational Divergence: It occurs when the sense conveyed by a single word is expressed by two or more words in one of the languages. For instance, Telugu uses 'snānaM ceyi' for 'to bathe' whereas it is expressed by 'kuḷi' in Tamil.

(24) Ta. *nāṉ kuḷi-pp-ēṉ.* Eng. I will bathe.
'I bath-fut-1p.sg'
Te. *nēnu snānaM cēs-tā-nu.*

'I bathing do-fut-1p.sg'

3.2.2. Categorical Divergence: Changes in category create categorical divergence. It is due to the mismatch between the Parts of Speech Categories of the words involved in the pair of languages considered. In Tamil *koṇṭu* is ambiguosly used as a postposition as well as a verb.

(25) Ta. *katti.y-aik* ***koṇṭu*** *aṟu-tt-ēṉ.* Eng. I cut it by knife.
'knife- inst **psp** cut-pst-1.sg'
Te. *katti-tō kōs-ā-nu.*
'knife-inst cut-pst-1.sg'

(26) Ta. *katti.y-ai.k* ***ko-ṇṭ-u*** *va-nt-ēṉ.* Eng. I came by taking the knife.
'knife-acc **hold-vpart** come-pst-2.sg'
Te. *katti-ni tīsukon-i vacc-ā-nu.*
'knife-acc take-vpart come-pst-2.sg'

3.2.3 Lexical Divergence: It arises when there is a lack of exact lexical equivalent but structure presents a translational equivalence between a language pair. Here, the literal translation of the source language word is substituted by a corresponding translational equivalent to resolve the problem. For instance,

(27) Ta. *eṉ-akku nīccal* ***teriyum****.* Eng. I know to swim.
'me-DAT swimming **know-fut-3p.sg.n'**.
Te. *nā-ku īta* ***vaccu****.*
'me-DAT swimming **come'**

(28) Ta. *avaḷukku* ***karppam*** *ēṟpa-ṭṭ-atu.* Eng. She became pregnant.
'she-dat **pregnancy** form-PRS-3p.sg.f'
Te. *āme-ku* ***kadupu*** *vacciMdi.*
'she-dat **belly** come-PST-3p.sg.n'

The divergences that are shown above are bridged by building a series of Transfer Grammar rules, Multi-word Expressions and Lexical Substitution module.

3. System Architecture and Handling Divergences: Current system is an assembly of various linguistic modules run on specific engines whose output is sequentially maneuvered and modified by a series of modules till the output is generated. The most crucial linguistic modules include, a Morphological Analyzer (MA), Parts of Speech Tagger (POS), Chunker, Simple Parser (SP), Multiword Expression Module, the Transfer Grammar Component (TG), a Lexical Transfer (LT) module consisting of a Conceptual Dictionary and a Bilingual Dictionary, an Agreement module (AGR) and a Morphological Generator (MG) besides a number of miner modules.

(1) Syntactic divergences are handled by Transfer Grammar Module where the structural transformations are carried out.
(2) Divergences relating to agreement are managed by a series of procedures reconstructing target language agreement.
(3) Lexico-semantic divergences are cleared by Lexical Transfer module which consist of a Conceptual (synset) dictionary and a stand-by Bilingual dictionary. In addition to these Multi Word Expressions (MWE) module involving a set of collocations.

The architecture of this system is based on analyze-transfer-generate paradigm. The flow of the input sentence in the system is given in fig:1. All the modules have been integrated on the dashboard, a tool, where the data flow in the pipeline is configured (ILILMT, 2007).

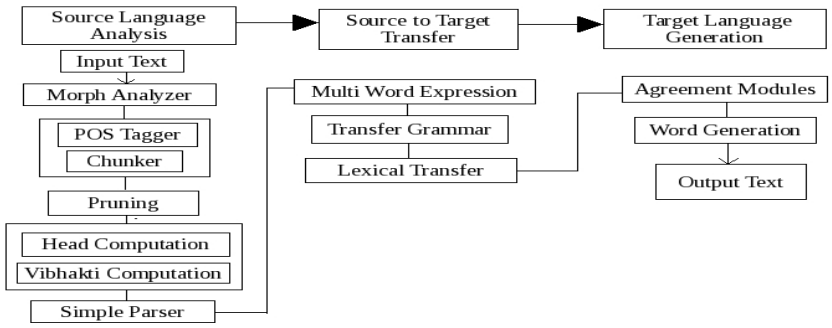

Fig:1 IL-IL MT Tamil-Telugu MT Architecture

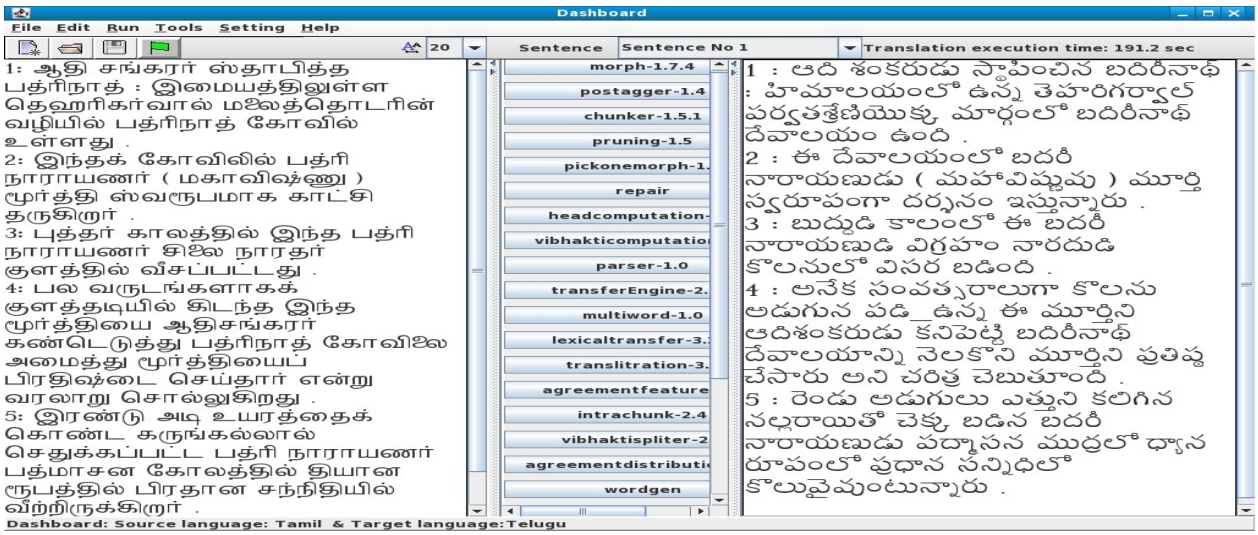

Fig:2 Sample Input and Output of Tamil-Telugu MT.

Conclusions: The system for translating between Tamil-Telugu (Bidirectional) is built and evaluated (IL-IL MT 0-4 scale evaluation, 2007) continuously in order to ensure enhanced quality output. It can be used to translate web pages or text material from books, magazines, newspapers etc. written in standard language.

References:

Comrie, B. 1991. *Form and Function in Identifying Cases*. In Paradigms: the Economy of Inflection. F. Plank (ed.), (Empirical Approaches to Language Typology 9), 41-55. Berlin : Mouton de Gruyter.

Dorr, Bonnie. 1993. *Machine Translation: A View from the Lexicon.* Cambridge, Mass: The MIT Press.

ILMT Consortium. 2007. *ILMT SRS and Functional Specifications (mimeo)*. Hyderabad.

Krishnamurti, Bh and Gwynn, J.P.L. 1985. *A Grammar of Modern Telugu*. New Delhi: OUP.

Lehmann, Thomas.S. 1989. *A Grammar of Modern Tamil.* Pondicherry: Pondicherry Institute of Linguistics and culture.

Subbarao, K.V. 2010. *South Asian Languages : A syntactic Typology.* Cambridge: *Cambridge University Press (in press)*.

Subbarao, K. V. and P. Bhaskararao. 2004. *Non-nominative Subjects in Telugu*, in P. Bhaskararao and K. V. Subbarao (eds.), *Non-nominative Subjects, vol. II.* (Amsterdam/Philadelphia: John Benjamins), 161–196.

Uma Maheshwar Rao, G. 2002. *A Computational Grammar of Telugu*. mimeo (350pages). Hyderabad: University of Hyderabad.