# Implementation of Transfer Grammar in Telugu - Hindi Machine Translation System

Christopher Mala
Center for Applied Linguistics and Translation Studies
University of Hyderabad
LTRC,IIIT-Hydearbad, Gachibowli
chirstopher.mpg08@research.iiit.ac.in

## Abstract

This paper describes experiments on Transformation of Grammar from one language to another while translating text through machine. It is known that every language has its own phenomena and its own way of representation. While translating text from one language to another it is very important to retrieve these language phenomena information of target language from source language, which may be absent in the source language. These language dependent phenomena can be seen alot while translating languages of two differnt language family. In this paper we have tried to explain how grammar is been transfered from Telugu (Dravidian language family) to Hindi (Indo-Aryan family).

## 1 Introduction

### 1.1 Transformational Grammar (TG) Definition

Transformational grammar seeks to identify rules (of transformation) that govern relations between Chunks of a sentence, on the assumption that there exists a fundamental structure beneath the word order of any language. Transformational grammar is the starting point for the tremendous growth to linguistic studies since 1950s.

### 1.2 Why Transformation Grammar is Required

The usual usage of the term 'transformation' in linguistics refers to a rule. For example, a typical transformation in TG is the operation of subject-auxiliary inversion (SAI). This rule takes as its input a declarative sentence with an auxiliary: *"John has eaten all the heirloom tomatoes"*, and transforms it into *"Has John eaten all the heirloom tomatoes?"*. These rules were stated as rules that held over strings of either terminals or constituent symbols or both. X NP AUX Y => X AUX NP Y (where NP = Noun Phrase and AUX = Auxiliary) Transformations are no longer structure changing operations at all, instead they add information to already existing trees by copying constituents. The earliest conceptions of transformations were that they were construction-specific devices. A different transformation of raised embedded subjects into main clause subject position in sentences and yet a third reordered arguments in the dative alternation. With the shift from rules to principles and constraints, these construction specific transformations are morphed into general rules. Generalized Transformations (GTs) take small structures which are either atomic or generated by other rules, and combine them.

### 1.3 Rules and Description

A formal Linguistic operation which enables two levels of structural representation, Dependency parsing and Phrase Structure, which contains sequence of terminals and non-terminals. Where as a Transformational Rule consisting of a sequence of symbols rewritten, as equivalent corresponding sequence to the source language. The input to Rule is the Structural Description, which defines the class of Phrase-Markers to which the rules can apply. The rule then operates a Structural Change on this input, by performing operations that were instructed in the rule.

Some of the changes made by the TG rules are given below:

1) *Transformation* (Movement) modifies an input structure by reordering the elements it contains. When this operation is seen as one of the moving elements to adjoin positions in a phrase-marker, it is known as Adjunction.

2) *Insertion* (Transformation) add new structure elements to the input sentence. Where as Deletion(Transformation) eliminates elements from the input sentence. etc..

Several models of transformation grammar have been presented since its first outline, that can manage some of the below listed functions.

a) Syntactic components b) Phonological Components c) Semantic components.

To design these grammar rule, we need to have strong knowledge about the source and the target languages. It is very important to understand the divergence between the two languages. Divergence at various levels like Lexical level, Morphological level and Syntactical level. Transformation Grammar(TG) deals with both Morphological and Syntactical divergence. TG is necessary in Translation to resolve the divergence between languages and produce translated text which is syntactically and semantically correct. Here we formulate few rules for the language that are of two different families.

Taking into consideration of the structural and semantic divergence of the both languages, it has been tried to formulate transfer rules for different sentence from Telugu to Hindi. In this we build rules by hypothesizing and then generalizing over them. These generalized rules represent contexts with constraints over semantic categories. We need to classify language divergence into various categories in different terms, all these divergence can be resolved by a set of TG rules. We can classify TG rules into Major and Minor. Some of them are:

- Copula
- Ergative
- Participles ("yA_huA","nA_vAlA")
- Conjuction (Ora)
- Modifying verb into Finite Verb
- Complementizer (-ani)
- Disjunction elements
- Discourse Markers

These are againe grouped into four and are explained briefly with examples in later half of the paper.

- Adding of Copula and other language specific data.
- Deletion of Grammar that is not required in the target language.
- Modification of the source language Grammar according to target language .
- Smoothing of the target language Grammar.

In this paper it has also been explained that Transfer Grammar engine which is of language independent and it can be used by training with rules. This study is being used in Indian Language - Indian Language Machine Translation project (IL-ILMT system) which is funded by Govt. of India (Minstery of Information Technology) being developed at CALTS lab in University of Hyderabad under the guidance of Prof. G. Uma Masheshwar Rao, Head, CALTS, HCU.

## 2 Introduction to Languages and their divergences

Telugu belongs to South-Central group (SD-II) of Dravidian languages. Morphologically Telugu is agglutinating in structure with no prefixes or infixes. Grammatical relations are expressed only by suffixation and compounding. Syntactically all Indian languages are of OV type, head-right-final and right-branching. The subject argument is generally expressed by a noun phrase (NP), but a post-position or case phrase with the nominal head in the dative case can also function as the subject, latter called as *'dative subject sentence'*. The predicate has either a verb or a nominal as head. Sentence with nominal predicate is equivalent sentence, which lack the copula or the verb 'to be' in Telugu. Nominal and verb predicates have different negative words which express sentence negation. A negation word is an inflected verb meaning 'to be' or 'to be not'. But this cannot be seen in Hindi, we can see the negative words as separate lexical items. Non-finite verbs, which head sub-ordinate clause, have affirmative and negative counter parts in Telugu . The arguments of NPs which occur as complements to a verb, are derive from the semantic structure of a verb; for instance, an intransitive verb require only one argument Agent/Object, where as transitive verb requires Agent+Object: a causative verb requires, Agent(causer) + Agent(casuse)+Instrument+Object. The passive voice is rarely used in modern Dravidian Languages.

## 3 How to use T.G in Machine Translation System

### 3.1 Flow of M.T

After analysing the input text of the source side. It has to be passed for lexical transfar. Before passing to lexical transfar, the process of transfar grammar should be done to reduce the language divergence. Then target language generation is done. As shown in the below fig.

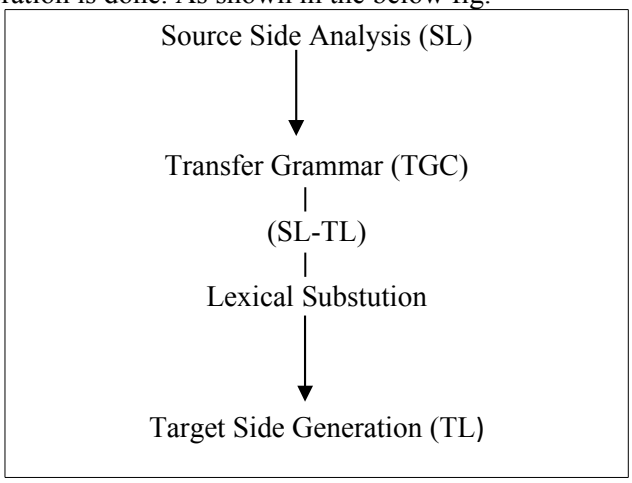


Fig 1: Structre of MT.

### 3.2 Transfer Grammar Rule Format Specifications

A grammar is a way to formally describe the structures of a language through a set of rules. Several formalisms have been developed for such descriptions in the field of NLP. PSG is a purely syntactic approach which uses a set of phrase structure rules to write the grammar of a language. It is

constituency based and the order of elements in a sentence is implicit in it. DG, on the other hand, tries to capture the semantic relations of the elements in a sentence.

For writing the transfer grammar rules a rule format needs to be specified. And since Indian languages are structurally very similar it is possible to achieve a high degree of correct transference without going to a deeper level of sentence analysis, i.e. a fully parsed sentence. Therefore, the transfer grammar format should also be able to handle shallow parsed inputs. For this level, the TG have rules that take chunks (for PSG) or bags (for DG) as inputs. For some special cases, a simple parsed (see below) level can also be accepted.

The rules would be stated differently in the PSG and DS formalisms. Conventions need to be defined for both these formalisms. However, before going into specifications of rules in a particular format it is important to identify the rule requirements. The transfer grammar rules would be stating the structural changes from the (Source Language) SL to (Target Language) TL. Rules would have an LHS and an RHS.

The format of a transfer grammar rule would have two parts – the Left Hand Side (LHS) part and the Right Hand Side (RHS) part. Therefore, the format of the rule is LHS => RHS

A Left Hand Side (LHS) and a Right Hand Side (RHS) which are separated by the symbol '=>'. The symbol '=>' stands for 'transfer to'. The LHS has the input from the source language – Telugu in this case and the RHS has the expected output of the rule for the target language. Therefore, the rule states that if the source language has a structure with two NPs in a sequence and they are related to each other by a genitive relation then a genitive marker should be inserted on the RHS. This is stated by changing the value of the attribute 'cm' from LHS (cm-UNDEF) to RHS (cm="*kI*").
Ex: NP~1(({<case="gen",cm="UNDEF">})) NP~2 => NP~1(({<case=gen,cm="*kI*">})) NP~2

## 4 Adding of target language specific data (Copula and ergator)

In this, data has handled, that is missing in the source language but it is very necessary in the target language to get proper translation. A few of the things are discussed below.

### 4.1 Handling of Obligatory Transformation

As it is known that the oblique form for common nouns in Telugu take "*ti*" as case maker (*oVMti, iMti*) for proper nouns its oblique form is "*du*" (*rAmudu*). But in Hindi there is only one case marker for oblique nouns (*kA*).

Rule: NP~1(({<case="o",tam="ti">})) NP~2 => NP~1(({<case="o",cm="kA">})) NP~2

### 4.2 "hE" insertion

Noun phrase (NP~1) is followed with an Adjective(NP~2) in source language (SL telugu), but in Hindi we need a copula in the target language at the end of the sentence.

Ex: (Tel) *rAmudu maMcivAdu*.

(HIN) *rAma accA vAlA* ***hE***.

The rule for the above example is given below:

Rule: NP~1 NP~2(({<lcat="adj">})) => NP~1 NP~2 +VGF(({hE%VM<root="hE", lcat="v", gen="m", num="sg", per="3", tam="hE">}))

#### 4.2.1 Example 2

If there is no verb in the source side then insert hE before the sentence ends.

Ex: (Tel) *rAma lakRamaNulu annaxammulu*.

(Hin) rAma lakRamana BahI hE.

Rule: VGF(({.%SYM})) => VGF((+{<root="*hE*",lcat="v",tam="*hE*">}{.%SYM}))

### 4.3 "Ora" insertion

If there are two noun phrases (NP~1,NP~2) with any long vowel as case marker then a conjunction is inserted in between these two noun phrases.

Ex: (Tel) *I waragawilo kurchIlu ballalu unnAyi*.

(Hin) *yaha kakRa me kursi Ora meja hE*.

Rule:NP~1(({<cm="*A*">})) NP~2(({<cm="*A*">})) => NP~1(({cm="0"})) +CCP(({*Ora* %CC<root="*Ora*",lcat="conj">})) NP~2(({<cm="*A*">}))

### 4.4 "ne" insertion

A direct noun phrase(NP~1) is followed with an oblique noun phrase(NP~2) and a verb phrase (VP) in the source side. Then in the target side "ne" is inserted in the first noun phrase (NP~1). And rest are retained.

Ex: (Tel) *rAmu puswakaM caxivAdu*.

(Hin) *rAma ne puswaka paDA*.

Rule: NP~1 (({<case="d",cm="">})) NP~2 (({<case="o">})) VGF(({<tam="*A*">})) => NP~1(({<case="o",cm="*ne*">})) NP~2(({<case="o">})) VGF(({<tam="*A*">}))

## 5 Deletion of Grammar that is not required in Target side

In this we are trying to frame rules to delete the information that is required in the target language from source language. Here are some of the examples that explain how deletion is done in Transformation Grammar.

### 5.1 Example:1

The word "*samayAnni*" in telugu will be having root as "*samayaM*" and case marker as "*ni*", but where as in hindi it is "*samaya*" with case marker as "0". So case marker is dropped in target side. Case marker "*ni*", and word ending with *$x.aM*, and lexical category "noun", can be dropped in target side.

Rule: NP(({<root="*samayaM*",cm="***ni***">})) => NP(({<root="*samayaM*",cm="***0***">}))

### 5.2 Example:2

If a 3rd person pronoun is having case marker as "*ki*" in source side, then it should be dropped in the target side.

Rule: NP(({<root="*ixi*",lcat="pn",cm="***ki***">}{*gAnu*%RP})) => NP(({<root="*isa*",cm="***0***">}{*gAnu* %RP\}))

## 6 Modifying in the Source side Grammar according to Target side

### 6.1 Example:1

In Telugu, any finite verb is ending with "*-ani*", example a verb like "*ceVppamani*" (*keha kara*) where "*ceVppu*" is the base form. The participle "*-ani*" means "*kara*". In Telugu we can see this "*-ani*" within the word, but Hindi "*kara*" is and aux-verb. So it has to be denoted as post position to the main verb (VM).

Rule:

VGF(({<tam="*$x.ani*">})) => VGF(({<tam="$x">})) +NP(({ani %PSP<root="*ani*",lcat="psp">}))

### 6.2 Example:2

In this example we can see a direction nominal which case marker is as "*na*" (nominative), this case marker is converted in locative marker "*lo*" in the target side.

Rule:

NP(({<root="*paScimaM*",lcat="n",cm="***na***">})) => NP(({<root="*paScimaM*",lcat="n",cm="***lo***">}))

another rule in which "***gA***" is converted to "***lo***".

NP(({<root="*BAgaM*",lcat="n",cm="***gA***">})) => NP(({<root="*BAgaM*",lcat="n",cm="***lo***">}))

### 6.3 Example:3

If any non-finite reduplicated verb is occurred in the sentence, eg: "*ceVppi ceVppi / cUsi cUsi / wini wini*", we even have the tense reduplication also. But in Hindi, we can see the tense reduplication is not possible. And the appropirate word for this reduplication verb is "*bawA bawA kara*".

So here one of the tense marker is dropped in the source side.

Rule:

VGNF~1(({<lcat="v",tam="i">})) VGNF~2(({<lcat="v",tam="***i***">})) => VGNF~1(({<tam="***0***">})) VGNF~2(({<tam="i">}))

## 7 Smoothing of the target language Grammar

In telugu, verb phrase like "*caMpina puli*", "*winina palYleVM*" etc. have a lot of ambiguity. Lets take example "*caMpina puli*" which mean "*mAra ne vAlA Sera*", have tam as "*-ina*" for the verb main in the verb group. It is know that for the tam "*-ina*", its corresponding hindi has two values, "*hE-jo-vaha*", "*wA-hE-jo-vaha*" and "*yA-hE-jo-vaha*". Depending on the context we have to choose the correct tam. When we have "*yA-hE-jo-vaha*" we should compress it to "*yA*", and when it is "*wA-hE-jo-vaha*" it is substuted with "*ne-vAlA*".

Example 1: VGF(({<lcat="v",tam="***wA-hE-jo-vaha***">})) => VGF(({<tam="***ne-vAlA***">}))

Example 2: VGF(({<lcat="v",tam="***yA-hE-jo-vaha***">})) => VGF(({<tam="***yA***">}))

## 8 Conclusion and Evaluation

### 8.1 Evaluation

A test is made to know human understandability of the text after implementaing the Tranfer Grammar in the IL-ILMT system.

For this test, 30 sentences in different construction are translated by MT system and given for manual evaluation. Their response is given below.

| Reader-1 | Rank 1-2 | Rank 3-5 | Accurecy |
|---|---|---|---|
| Before (TG) | 14 | 16 | 53.33 |
| After (TG) | 09 | 21 | 70.00 |

Table 1: Data-Sheet of Reader-1

Scale of Ranking is given as:

Rank 1 = Non Sense , Rank 2 = Not understandable, need major change , Rank 3 = Understandable but need minor change , Rank 4 = Good but very minor change for perfect and Rank 5 = Perfect Translation.

### 8.2 Conclusion

As we know that every language has its own phenomena which is called as language divergence. These language divergences should be taken care while translating the text form from one language to another. This is should be even more carefully handle when the translation is between cross language families. These divergence can be of **Lexical level, Morphological level and Syntactic level**. Transformation Grammar (TG) deals with both Morphological and Syntactic divergences. *TG is plays a vital reduce these divergence while Translating text and increase the understanbility of the reader.*

## References


Beames, John. 1966. *A Comparative Grammar of Moder Aryan Languages of India*. Delhi.

Gildea, Spike. 2000. *Reconstructing Grammar: comparative, linguistics and grammaticalization*. John Benjamins Publishing Company.

Masica, Colin Paul. 1996. *The Indo-Aryan Languages*. Cambridge University PressCambridge,UK.

Krishnamurti Bh. 2003. *The Dravidian Languages*. Cambridge University Press, Cambridge, UK.

Krishnamurti Bh and Gwynn J.P.L. 2003. *A Grammar of Modern Telugu*. Oxford University Press, Delhi, India.

Sjoberg, Andree F. 1992. *The Impact of the Dravidian on Indo-Aryan: An Overview* .

Robert Coldwell. 1891. *Comparative Grammar*.