J Indian Inst. Sci., Mar.–Apr. 1990, **70**, 131–143.

Analysis of case-endings in Indian languages and bidirectional machine translation between Indian languages and English

R. SANTHI SEELA* AND G. KRISHNA
Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India.

Received on January 9, 1989; Revised on December 15, 1989.

Abstract

This paper describes the analysis of case-endings in Indian languages and bidirectional machine translation between Indian languages and English, keeping the vocabulary to the required optimum and restricting the sentence syntactic structures to functionally viable patterns. The parser employed for syntactic checking adapts non-deterministic, backtracking, top-down ATN (augmented transition network) technique. The source statements are split into syntactically meaningful groups of words which are translated by employing the morphological rules of source and target languages. Finally the translated groups are transpositioned according to the word order of target language to get the translated version.

Key words: Case-endings, Indian languages, bidirectional machine translation.

1. Introduction

Work on machine translation has now spanned almost four decades. The great effort put into this field is largely because of the need to derive contextual meaning of the natural language sentences amidst the word ambiguities and syntactic structure complexities and the necessity of finding a systematic methodology for the machine to handle the tasks which are inherently persisting in human languages.

In India, the problem of communication among the people of different states is of vast proportions, as each state uses its own regional language. The diversity in Indian languages makes it impossible for Indians from different parts of the country to communicate with each other, unless both learn a common language. In a country where there are more than 15 regional languages and over 300 dialects, machine translation from an Indian language to English and *vice versa* and from one Indian language to another is of great relevance, especially for the common masses who depend mainly upon central information systems which are vital for decision making. *e.g.* fishermen and farmers whose activities depend on

*Present address: Flight Simulation, Aeronautical Development Establishment, Defence Research and Development Organisation, Bangalore 560 012, India.

weather forecasts, etc. Several efforts have been made elsewhere in the world, but in India machine translation is still in its infancy. This paper describes a system for translation from an Indian language to English and *vice versa* and from one Indian language to another.

The actual scripts of Indian languages are not used in this work, instead romanized scripts are employed[1]. With minor modifications, these scripts can easily be converted to original language scripts.

A simplified and specific set of rules were laid for the chosen natural language to enable a computer to process various tasks such as machine translation, man-machine interface, computer-assisted language teaching, automatic abstracting and so on. This paper discusses the analysis of case-endings in Indian languages in some detail, besides providing procedures for language translation using Tamil as reference language.

2. Analysis of case-endings in Indian languages[2]

Nouns, pronouns and verbs in Indian languages belong to inflexible categories. Adjectives are inflected when used predicatively but not when used attributively. Gender is a semantic-cum-grammatical category. Verbs are conjugated for tense, person, gender and number and they keep concord with the subject. There are, however, many verbal forms suffixed with particles which are impersonal and have no concord. Syntactically Indian languages also have the subject-object, complement-verb order with adjuncts generally preceding the qualified or modified items.

2.1. *Morphology of Indian languages*

Morphology is the study of morphemes and their arrangements in forming words. Morphemes are the minimal meaningful units which may constitute words or parts of words. The formal relationship of morphemes to each other is structural and positional, the structural relationship of morphemes includes three different morphemic types. These represent three basic morphological processes: (i) addition, (ii) replacement, and (iii) subtraction.

The words in Indian languages are found in sets of related forms. These sets fall into ten patterns, namely 1. noun, 2. numeral, 3. pronoun, 4. adjective, 5. adverb, 6. verb, 7. participle, 8. imitative, 9. echo word, and 10. interjection. The phrases obtained by adding inflectional suffixes to the root or stem show certain structural parallelism and allomorphs. Verb stems are classified into different sub-classes on the basis of the tense suffixes. For example, in Tamil, lexicon has grouped the verb under thirteen conjugational classes out of which one is the irregular class (Table I).

2.2. *Notes on structural relationship of morphemes*[1]

1. Stem ending LX is replaced by NX before adding the morpheme TX.
2. Stem ending L is replaced by N before adding the morpheme RH.
3. The last vowel U is subtracted from stem before adding the morpheme INH.
4. The last vowel U is subtracted and the consonant preceding it is reduplicated.
5. The last consonant LX is subtracted and TX is added before adding the tense marker.

Table I
Conjugational classes for Tamil language

Class no.	Present marker	Future marker	Past marker	Root verb (example)
1.	KIRH	V	T	*CEY*
2.	KIRH	V	TX (1)	*AALX*
3.	KIRH	V	RH (2)	*CEL*
4.	KIRH	V	RT	*ARHI*
5.	KIRH	V	INH (3)	*PANXNXU*
6.	KIRH	V	TX, TXTX (4) RH, RHRH K, KK	*NATXU*
7.	KIRH	P	TX	*UNX*
8.	KIRH	P	RH	*TINH*
9	KIRH (5)	P (5)	TX (5)	*KEELX*
10.	KIRH (6)	P (6)	RH (6)	*KAL*
11.	K KIRH	PP	TT	*PATXI*
12.	K KIRH	PP	NT	*IRU*
13	Irregular			

The numbers in parentheses refer to structural relationship of morphemes.

6. The last consonant LX is subtracted and RH is added before adding the tense marker.

Following are the rules for the formation of the future tense third person, neuter gender, singular and plural verbs.

1. Stem + UM for classes 1, 5 and 6
2. Stem + last consonant + UM for classes 2, 3, 7 and 8.
3. Stem + YUM for class 4
4. Stem + KUM for classes 9 and 10
5. Stem + KKUM for classes 11 and 12.

In all the above cases, person, gender, number endings are not appended. There are eleven personal endings (Table II).

Table II
Personal endings Tamil language

Sl no	Personal endings	Person	Gender	Number
1.	*Eenh*	I	Common	Singular
2.	*Oom*	I	Common	Plural
3.	*Aay*	II	Common	Singular
4.	*Iir*	II	Common	Honorific
5.	*Iirkalx*	III	Common	Singular
6.	*Aanh*	III	Masculine	Singular
7.	*Aalx*	III	Feminine	Singular
8.	*Aar*	III	Common	Honorific
9.	*Aarkalx*	III	Common	Plural
10.	*Atu*	III	Neuter	Singular
11.	*Anha*	III	Neuter	Plural

Indian language verbs are formed by adding the ordered pair of peripheral constituents which are mutually obligatory to the root. The two peripheral constituents are tense-marker and the PGN (person, gender, number) ending. Plural paradigm consists of three components, *viz.*, root, inflectional increment and plural suffix. The last component is static and does not change while the second component changes according to the noun. Plurals of different types of nouns in Tamil are given below.

Sl no	Singular	Root	Increment	Plural suffix
1.	*Maram*→ (Tree)	*Mara* +	*ng* +	*kal*
2.	*Puu*→ (Flower)	*Puu* +	*k* +	*kal*
3.	*Pal*→ (Tooth)	*Pa* +	*r* +	*kal*
4.	*Kappal*→ (Ship)	*Kappal* +	— +	*kal*
5.	*Eli*→ (Rat)	*Eli* +	— +	*kal*

Prepositional phrases are formed by adding preposition endings after adding the preposition jointer to the root. For some phrases preposition jointer varies according to the gender inflected by the root. Some examples of Tamil preposition phrases are given in Table III.

Table III
Prepositional phrases in Tamil

Preposition	Root gender	P-jointer
Iruntu	Neuter	il
Iruntu	Non-neuter	itxam
Ku	Neuter	ik
Ku	Non-neuter	irh
Aaka	Neuter	ikk
Aaka	Non-neuter	irh
.	.	.
.	.	.
.	.	.

2.3. *Case of homonymy and ambiguity*[4]

The problems of homonymy and ambiguity in a natural language are resolved mostly using the contexts of usage. To a certain extent it can be done by collocations. For example, in Tamil the word *pati* would mean 'to read', 'measuring container', and 'steps'. Each one of these meanings pertains to a proper context. The meaning 'measuring container' is always accompanied by the words like *ala* 'to measure', *vaangku* 'buy', etc., and the meaning 'step' is

always accompanied by the word like *natxa* 'walk', *eeru* 'climb', etc. By identifying these words and providing them in the dictionary with the indication of homonym the problem can be solved. However, identifying such collocational words for all the homonyms of a language is a difficult problem. With regard to the structural ambiguity it is even difficult to provide a solution through conventional approach of natural language processing (NLP) since it is not possible to instruct the computer with the proper context of occurrence. Consider the phrase, *'putiya puttakak katxai'* 'new book shop', it can not be understood from the phrase as to which noun the adjective *'putiya'* 'new' modifies. That is the two meanings *viz.*, 'fresh book shop' and 'new book shop' as opposed to 'old book shop'. The problems of natural languages such as syntactic and lexical ambiguity can be resolved in an NLP through proper means of parsing and knowledge representation techniques. Proper inferencing algorithms would help to identify and solve the problems of ambiguity in an NLP.

In Indian languages, at morphological level, the inflections of nouns, verbs, adjectives and adverbs have to be analysed and a specific set of rules have to be framed in order to produce the correct paradigms of the classes which will be used for the process of translation. As far as the syntax is concerned word order, co-references, etc., have to be in a simplified manner.

3. Syntactic analysis

Since translation is done phrase-by-phrase, it is necessary to check whether the input sentence is syntactically correct. The parser dealt here is non-deterministic backtracking top-down augmented transition network parser[5].

3.1. *Dictionary and morphological units*

The parser can derive the association between words and morphological units from the dictionary. The dictionary includes morphological units of the main word like part-of-speech, gender, number, person, tense, target equivalents, etc. Individual words often encountered contain different suffixes. In Indian languages, the verb is a combination of root word, tense marker, and PGN ending. Conjugated verbs are derived instead of explicitly giving separate entities in the dictionary. The parser uses explicit knowledge of the structure of words and it figures out whether the word is a variant of the one that is already in the dictionary.

So, it is necessary to have a pre-processor module which will extract the root verb from the verb pharse by identifying and removing the other morphemes which indicate case endings, tense, etc. In addition, it extracts the root of the noun even if it includes a plural ending. If noun phrase is a compound word containing preposition, then the word is decomposed into root + preposition.

3.2. *Transformational grammar*

The pre-processor module is called as transformational grammar unit. The transition

network is augmented to handle transformational rules to do the following tasks.

1. Remembering what already appeared in the sentence.
2. Manipulating morphological units on constituents.
3. Adding and deleting constituents.

To do these tasks, transformational rules perform

1. Put features on constituents of the sentence,
2. Move constituents,
3. Add constituents,
4. Delete constituents.

These transformational rules convert the deep structure of input sentence into a surface structure on which ATN grammar can be applied to parse the sentence, according to the rules given below.

1. *Subject-verb agreement*: Subject-verb agreement is a standard test for rule formalism in Indian languages. This informal rule checks the number and gender feature of subject and verb should match for agreement.

2. *Splitting prepositional phrases*: Prepositional phrase (PP) is a combination of modified noun phrase (NP) and preposition. This rule splits the PP and extracts the root NP and preposition.

For example,

Marattiliruntu $\longrightarrow$ *Maram* + *Iruntu*
(from the tree) $_{(t-r)}$ (tree) (from)

3. *Imperatives*: In Indian languages some imperative sentences do not have subject. This rule analyses verb phrase (VP) and on the basis of the gender reflected by VP, it adds the proper subject.

e.g:

Poo $\longrightarrow$ *Nii* *Poo* (gender: not honorific)
(go) $_{(t-r)}$ (you) (go)

Poongkalx $\longrightarrow$ *Niingkalx* *Poongkalx* (gender: honorific)
(go) $_{(t-r)}$ (you) (go)

4. *Question-element movement*: Yes-No question type sentences always have question-element inside the VP or NP. This rule splits the question element from the phrase and puts it at the end of the sentence.

e.g:

Ramana cenhrhaanh? $\rightarrow$ *Raman cenhrhaanh as?*
Raman cenhrhaanhaa? $\rightarrow$ *Raman cenhrhaanh aa?*
(Did Raman go?)

5. *Negative-element movement*: In Indian languages sentence has negative element inside the VP or NP. This rule extracts the negative element and puts it at the end of the sentence and modifies its phrase.

e.g.:

Raman cellavillai → *Raman cel illai*
(Raman did not go)

3.3. *Parsing technique*[6]

Top-down parsers try to match the grammar rules against the input, starting at the top-most rewrite rule (which usually involves the start symbol or sentence symbol S) and recursively moving towards the lower more specific rewrite rules. The parser is successful if a sentence can be constructed that matches the input sentence.

Top-down parsers are easy to write and modify. Rules that are more likely to be used can easily be placed ahead of the less-likely rules by enhancing performance. Top-down parsers can be slow. If all the rules at a particular level fail, the parser backtracks up to the previous level to try another rule. During backtracking the same constituents may be analysed many times. Augmented transition networks are very powerful. They have the following features: (i) registers which can store conditions or information on a global basis regardless of which particular subnetwork is being processed, (ii) conditions which allow arcs to be selected if registers indicated certain conditions, and (iii) actions which allow arcs to modify the structure of data.

Note that the arcs in an ATN can be labelled, not only with words, word classes and non-terminals, but also with arbitrary tests that depend on the state of the global registers. The mood of the sentence is analysed by the parser. The parser assigns the mood of the sentence in the variable mood. Similarly the voice is analysed and assigned to the variable voice. If the sentence is syntactically wrong, the translation will be aborted. Otherwise, it proceeds to the translation phase.

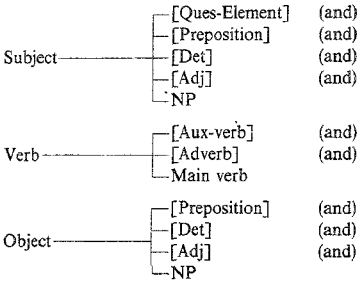
4. Automatic demarcation of SVO (subject verb object) group

The words of the sentence are grouped into subject, verb and object. The format of these in Indian languages is given below.

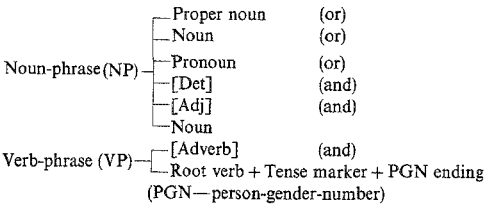
Subject	[Ques-Elem]	(and)
	[Det]	(and)
	[Adj]	(and)
	[NP]	(and)
	[PP]	
Verb	[Adverb]	(and)
	Verb	

Subject and object groups should have at least one noun-phrase (NP) or prepositional-phrase (PP).

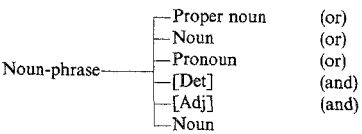
Format of these groups for English is given below.



These groups are further divided on the basis of phrases. The format of those for Indian languages,



Prepositional phrase (PP) → Modified NP + preposition ending.
In English,



Verb-phrase → [Aux-verb] (and)
[Adverb] (and)
Main verb

Prepositional-phrase → Preposition (and)
NP.

At the end of this process the groups contain phrases in order.

Subject → (⟨S – P1⟩⟨S – P2⟩...⟨S – PN⟩)
Object → (⟨S – P1⟩⟨S – P2⟩...⟨S – PN⟩)
Verb → (⟨S – P⟩)
S – P—Source-phrase.

Each phrase is analysed by following the morphological rules of the source language and morphological units are attributed a value in order to generate the sentence. These codified values enable the analysis of the sentence. Each language has its own morphological rules.

For example, Tamil verb-phrase *Atxittaanh* is analysed as,

Atxittaanh	⟶	⟨*Atxi*⟩	⟨*tt*⟩	⟨*aanh*⟩
(Hit)		(Root)	(T – M)	(PGN)

Root-word → *Atxi*
Tense → Past tense
Gender → Masculine
Person–number → 3–Singular.

In exceptional cases this root-word is further modified to get the correct one.

For example, *Marattiliruntu* is analysed as,

Marattiliruntu	⟶	⟨*Maratt*⟩	⟨*il*⟩	⟨*iruntu*⟩
(From the tree)	M-rules	⟨Root⟩	⟨P-J⟩	⟨Preposition⟩

Maratt ⟶ *Maram.*
(Exceptional rules)

Root ⟶ *Maram* (Tree)
Preposition → *Iruntu* (From)

After analysing the sentence the process proceeds to the phase of translation.

5. Translation of syntactic groups

The root of the phrases is referred in the dictionary and their target equivalents are extracted. Morphological units are added by following the morphological rules of the target language, target phrases are constructed according to the morphological units of the source

language which were set during the analysis of the source language phrases. Translation of every group is then generated.

For example, Tamil VP *Atxittaanh* is translated in Kannada as,

S–Root → *Atxi*————*Hode*——(Target root)
S–Tense → Past———'d'——(Past tense marker in target language)
S–Gender → Masculine
S–Number–person → 3 Singular——*anu* (PGN ending in target)
'*Atxittanh*' (Source)————'*Hodedanu*' (Target VP)
(Tamil)

In exceptional cases, the target root-word is modified further to the correct one. After generating the translated phrases, it is necessary to reverse the order inside the groups to avoid the ambiguity occurring in certain types of sentences.

For example, in translation from English to Tamil,

Source—'The cat on the table ate the rat'
S–Sub—(⟨The cat⟩⟨on the table⟩)
S–Obj—(⟨the rat⟩)
S–Verb—(⟨ate⟩).

Translating phrase-by-phrase, the target groups (Tamil) are,

T–Sub—(⟨*Poonai*⟩⟨*Meesaiyinmeelirunta*⟩)
T–Obj—(⟨*Eliyai*⟩)
T–Verb—(⟨*Tinhrhatu*⟩).

Translated sentence is, *Poonai meesaiyinhmeelirunta eliyai tinhrhatu.* Here, the translated sentence means that 'The cat ate the rat on the table'. By applying phrase-reversing rule in the target groups, this ambiguity is removed.

Phrase-reversing rule:

S–Group—(⟨S–P1⟩⟨S–P2⟩...⟨S–PN⟩)
T–Group—(⟨T–P1⟩⟨T–P2⟩...⟨T–GN⟩)

After applying this rule in the above example, the resultant target groups are,

T–Sub—(⟨*Meesaiyinhmeelirunta*⟩⟨*Poonai*⟩)
T–Obj—(⟨*Eliyai*⟩)
T–Verb—(⟨*Tinhrhatu*⟩).

After arranging the groups, target sentence *Meesaiyinhmeelirunta poonai eliyai tinhrhatu* gives the correct meaning.

6. Target transposition

This deals with the rearrangement of the groups to form target sentence. The word-order pattern of Indian languages is subject, object, and verb (SOV) and that of English is subject,

verb and object (SVO). In the process of translation of English to Indian languages, if the verb is 'BE' type (contains only weak verbs), the translated sentence contains only subject and object. If the verb is of ordinary type the sentence is presented in the order of subject, object and verb. The translated target groups are finally arranged according to the language's word order to form the target sentence (fig. 1).

7. Sample output

Source (English) : The gift was given by Kannan to Sudha
Target (Tamil) : *Parisu sutxaavukku kanhnhanaal kotxtxukkap patxtxatu.*

Source (Tamil) : *Raama siitayai atxittaanh*
Target (Kannada): *Raama sitege hodedanu.*

Source (Kannada): *Meri jaannige pustakavannu Kottalu*
Target (Hindi) : *Meryne jaanko kitab dee.*

Source (English) : The very fat boy slowly put a beautiful toy on the big table.

1. Target (Tamil) : *Mika kuntxu paiyanh periya meesaiyinhmeel oru alakaanha pommaiyai metuvaaka vaittaanh.*

2. Target (Kannada) : *Anti dappa huduganu dodda mejinamele ondu sundara aatikeyannu nidhanvagi ittanu.*

3. Target (Hindi) : *Bahut mote ladakene bade mejpar paryek sundara khilona aahiste rakha.*

Source (Tamil) : *Kamala periya marattiliruntu kutittaalx.*
Target (English) : Kamala jumped from the big tree.

8. Conclusion

This paper has its own limitation even with the selected vocabulary and syntactic patterns, a global translation can not be achieved for the following reasons. Language semantics are not taken into consideration fully and threadbare test for all combinations of syntactic were not performed. It is realised that languages are different, not only in how they say things but in what they say. At times, translation becomes necessarily vague, involving loss of information.

The efficiency of the translation here depends on the efficiency of the 'parser' and the size of the dictionary. Though the machine works as per the set of rules framed by the operator, the translation sometimes lacks the delicacy and smoothness of the spoken target language and idiomatic expression of the source language. This work can further be augmented with further refinements for man-machine interfacing, translation of languages, computer-assisted language teaching, etc.

SOURCE (ENGLISH) : THE GIFT WAS GIVEN BY KANNAN TO SUDHA

TARGET (TAMIL) : பரிசு சுதாவுக்கு கண்ணனால் கொடுக்கப்பட்டது.

SOURCE (TAMIL) : ராமா சீதையை அடித்தான்.

TARGET (KANNADA) : ರಾಮನು ಸೀತೆಯನ್ನು ಹೊಡೆದನು.

SOURCE (KANNADA) : ಮೇರಿಯು ಜಾನ್‌ನಿಗೆ ಪುಸ್ತಕವನ್ನು ಕೊಟ್ಟಳು.

TARGET (HINDI) : मेरी ने जान को किताब दी ।

SOURCE (ENGLISH) : THE VERY FAT BOY SLOWLY PUT A BEAUTIFUL TOY ON THE BIG TABLE.

TARGET (TAMIL) : மிக குண்ட பையன் பெரிய மேசையின்மேல் ஒரு அழகான பொம்மையை மெதுவாக வைத்தான்.

TARGET (KANNADA) : ಆ ತಿ ದಪ್ಪ ಹುಡುಗನು ದೊಡ್ಡ ಮೇಜಿನ ಮೇಲೆ ಒಂದು ಸುಂದರವಾದ ಆಟಿಕೆಯನ್ನು ನಿಧಾನವಾಗಿ ಇಟ್ಟನು.

TARGET (HINDI) : बहुत मोटे लडके ने बडे मेज पर एक सुंदर खिलौना आहिस्ते रखा ।

SOURCE (TAMIL) : கமலா பெரிய மரத்திலிருந்து குதித்தாள்.

TARGET (ENGLISH) : KAMALA JUMPED FROM THE BIG TREE.

FIG. 1. Translated output in Indian scripts.

References

1. SANTHI SEELA, R. AND KRISHNA, G. — Machine translation from English to Tamil, *First National Conf. on Knowledge-Based Computer Systems*, Bangalore, India, 1988, pp. 53–55.
2. SANTHI SEELA, R. AND KRISHNA, G. — Morphological analysis and bidirectional machine translation (with special reference to Indian languages), *Regional Workshop on Computer Processing of Asian Languages*, Asian Institute of Technology, Bangkok 10501, Thailand, 1989.
3. NAMPERUMAL, S. — *Application of programming languages for processing natural languages with specal reference to English and Tamil*, Ph.D Thesis, Indian Institute of science, Bangalore, India, 1981.
4. CHELLAMUTHU, K. C AND VASU, R. — Some selected problems in natural language processing, *First National Conf. on Knowledge-Based Computer Systems*, Bangalore, India, 1988, pp. 56–63.
5. CHARNIAK, E. AND MCDERMOTT, D. — *Introduction to artificial intelligence*, 1985, Addison-Wesley.
6. WOODS, W. A. — Transition network grammars for natural language analysis, *Commun. ACM*, 1980, **13**, 591–606.