A System for Syntactic Structure Transfer from Malayalam to English

Latha R Nair
Cochin University of Science and Technology

David Peter S
Cochin University of Science and Technology

Renjith P Ravindran
Cochin University of Science and Technology

ABSTRACT
This paper describes the design and development of a system for syntactic structure transfer of Malayalam sentences to English. A syntactic structure transfer module is required in machine translation systems using a transfer based approach. The system uses a rule based approach. It makes use of rules of morphology of both Malayalam and English and syntactic structure transfer rules between Malayalam and English sentences. The rules of morphology and sentence syntax are used by the shallow parser in the creation of the parse tree for the source language sentence. The transfer rules are used for generating target structure. The system uses a bilingual dictionary consisting of words in the source and target language along with their lexical category. Regular expression notation is used for the representation of language syntax. The parsing algorithm is a general one and the system can be used to handle other language pairs by changing the set of rules.The results obtained are encouraging and the work can be extended to the creation of a full fledged machine translator from any Dravidian language to English since all of these languages exhibit structural homogeneity.

General Terms
Transfer based machine translation.

Keywords
Transfer based Machine Translation, Parser, Malayalam language, Syntactic Reordering, Regular Expressions.

1. INTRODUCTION
Machine processing of Natural (Human) Languages has a long tradition, benefiting from decades of manual and semi-automatic analysis by linguists, sociologists, psychologists and computer scientists among others. One of the central design questions in machine translation(MT) is the syntactic structural transfer, which is the conversion from a syntactic analysis structure of the source language to the structure of the target language. This paper describes a system for syntactic structure transfer from Malayalam, a Dravidian language popular in South India, to English. This system takes a paragraph of Malayalam sentences as input and produces equivalent English sentences after reordering. The system uses a morphological parser for context sense disambiguation and chunking, a syntactic structure transfer module and a bilingual dictionary. This system does not rely on a stochastic approach. Rather, it is based on a rule-based architecture along with various linguistic knowledge components of both Malayalam and English. The system uses two sets of rules: rules for Malayalam morphology and rules for syntactic structure transfer from Malayalam to English. The rules are given in regular expression form. The system is based on artificial intelligence techniques. As the rules are separately stored the system can be extended to any level.

2. PREVIOUS WORK
The English to Hindi MT system Mantra, developed by Applied Artificial Intelligence (AAI) group of CDAC, Bangalore, in 1999 uses transfer based approach. The system translates domain specific documents in the field of personal administration; specifically gazette notifications, office orders, office memorandums and circulars. It is based on lexicalized tree adjoining grammar (LTAG) to represent English and Hindi grammar which are used to parse source English sentences and for structural transfer from English to Hindi. It uses preprocessing tools like phrase marker, named entity recognizer, spell and grammatical checker. It uses Earley's style bottom up parsing algorithm for parsing. The system provides online addition of grammar rule. The system produces multiple translation results in the case of multiple correct parses.

An English to Kannada MT system is developed at Resource centre for Indian Language Technology Solutions (RC_ILTS), University of Hyderabad by Narayan Murthy et.al.. This also uses a transfer based approach and it can be applied to the domain of government circulars. The project is funded by Karnataka government. This system uses Universal Clause Structure Grammar (UCSG) formalism [1]. The technique is applied to English_ Telugu translation as well.

A hybrid approach has been tried at AU-KBC Research centre, Anna University for syntactic transfer in a Tamil to Hindi MT system [2]. It uses CRFs for identifying the clause boundaries in the source language, transformation based learning for extracting the rules and semantic classification of postpositions for choosing semantically appropriate structure in constructions where there are one to many mapping in the target language.

3. SYNTACTIC STRUCTURE TRANSFER SYSTEM
The basic building blocks of the system are shown in figure 1. The system has the following modules:

i). morphological parser for context sense disambiguation and chunking

ii). The transfer module transfers the source language structure representation to a target language representation.

The grammar rules for Malayalam and some of the transfer rules for transferring source parse tree to target parse tree are stored in two separate files. Some of the transfer rules are embedded in the source code.

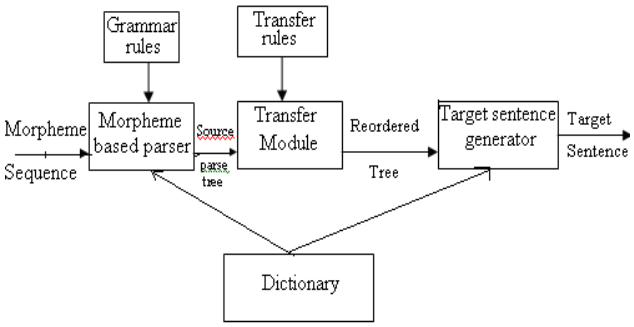


Figure 1.Block diagram

3.1 Syntax of Malayalam Language

Malayalam is mostly an S-O-V language. The default or unmarked order of constituents is Subject first, then the Object and finally the verb. However, Malayalam, being a relatively free word order language, permits substantial amount of freedom in the order of constituents although normally the verb remains in the sentence final position. Word order becomes less important mainly because noun groups are marked for cases and the verb agrees with the subject in gender, number and person. In fact, subjects and objects are often dropped. Normally all modifiers precede the modified. There are a variety of subordinate clauses. Subordinate clauses also precede the main clause. They typically involve special non-finite forms of verbs which occur invariably in the clause final position and mark the right hand boundary of the respective clauses. All these assertions are taken as rules. There are exceptional situations where deviations from these rules are possible. Also, most of these rules apply not only to Malayalam but to Dravidian languages in general. The subject of a sentence is expressed by a noun group in the nominative case. The nominative case marker is empty. Malayalam also permits dative subject constructions where the understood subject is indicated by a noun group in dative case whereas the surface subject appears in the nominative case.

3.2Parser Module

The input to the parser is a sequence of morphemes in the sentence. An example is shown below:

Malayalam sentence :
Malayalam sentence:സീതയുടെപൂച്ചയൊരെലിയെത്തിന്നു- (1)

English sentence: Seetha's cat ate a rat.

The morpheme sequence for the above sentence is:

സീത ഉടെ പൂച്ച ഒരു എലി എ തിന്നു--(2)

morpheme based translation into English for 2 is:

Seetha 's cat a mouse ate

Parser does the following tasks : 1) It groups the input sequence of morphemes into chunks [3, 4]. 2) Word sense disambiguation based on morpheme tags [5]. The parser uses a depth first approach with backtracking [6, 7, 8]. The output of the parser is a parse tree for the next module. The parser uses the syntax rules for the morpheme sequences in Malayalam sentences in regular expression notation. A set of syntax rules in the regular expression form are given below:

1. S-> NP*VP

2. NP-> ADJG* N | ADJG* N NA | ADJG* N PL NA | NPC

3. VP ->ADV* V VA| INF VG

The first rule says that a simple sentence is a sequence of noun chunks followed by a verb chink. According to second rule, a noun phrase chunk consists of a set of adjectives followed by a noun which is followed by plural or case suffix or a set of nouns with conjunction. The third rule says that a verb chunk consist of a sequence of adverbs followed by a verb and verbal suffix or an infinitive followed by a verb group. Rules for each of the chunks on the right hand side is recursively defined in the same way. The chunks were chosen in such a way that they form a subtree to be reordered in the source parse tree to get the target parse tree. Only a subset of the derived set of rules is shown above.

The sequence of morphemes for another sample Malayalam sentence is given below. The parse tree generated for the sentence using the grammar rules is shown in Figure 2(a).

Input set of morphemes: "രാമൻ", "രാവണൻ", "എ", "കൊന്നു", "അപ്പോൾ", "സീത", "സന്തോഷിച്ചു"

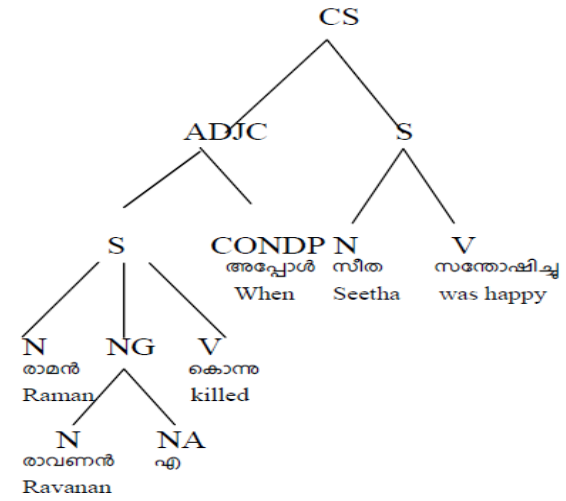
English words for the input sequence: Raman Ravanan killed when Seetha was happy

Correct English translation: Seetha was happy when Raman killed Ravanan.

The above sentence is a complex sentence with one adverbal clause and one principal clause. The adverbal clause contains a sentence followed by the suffix അപ്പോള് (when) with the morpheme tag CONDP which marks the end of the clause. The sentence contains a subject, object and a verb. The principal clause contains a subject and a verb. The hierarchical chunks of the sentence are:

CS(ADJC(S(N(രാമൻ),NG(രാവണൻ,എ),V(കൊന്നു)),CONDP (അപ്പോൾ)) , S(N(സീത),V(സന്തോഷിച്ചു))

CS(ADJC(S(N(Raman),NG(Ravanan,),V(killed)),
CONDP(when)),S(N(Seetha),V(washappy))



(a)

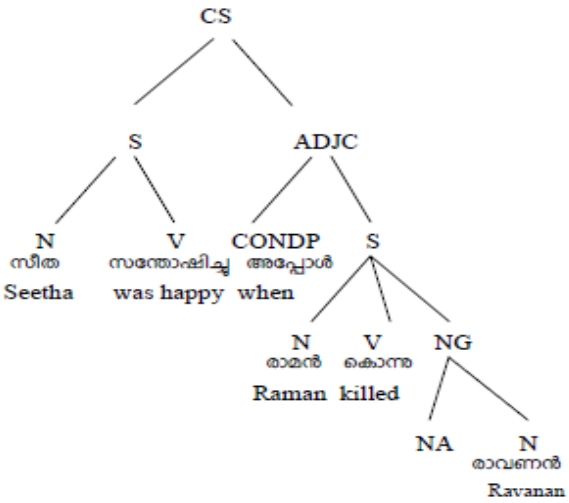


Figure 2. Parse tree before and after reordering

3.3 Syntactic Structure Transfer Module

Malayalam and English belong to two language families and their sentence structure differs a lot. So a simple morpheme to morpheme mapping will not give the correct translation for the input sentence. This is clear from the above example.

The transfer module developed transfers the source language structure representation to a target language representation. This module needs the subtree rearrangement rules by which the source language sentence syntax tree can be transformed into target language sentence syntax tree. The system performs most of the commonly needed reordering for Malayalam to English translation.

The system uses a set of transfer rules written in the syntax shown in Table 1.

Table 1. Transfer rules

	Malayalam structure	English structure
1	PP-> NP P	PP->P NP
2	VG->ADV V	VG->V ADV

The first rule says that the order of case suffix and noun chunk should be interchanged in a prepositional chunk. Second rule says that in verbal chunk the adverb and verb should be interchanged.

The tree after reordering for the above Malayalam sentence using the transfer grammar rules identified is shown in Figure 2(b). It uses three reordering rules:

i). The noun and preposition are exchanged in a noun group. Thus the group (NG(N(രാവണൻ), NA(എ)) is changed to NG(NA( ), N(രാവണൻ)).

ii). The subtree for sentence and adjectival clause suffix are exchanged in the subtree for adjectival clause.

iii). The subtrees for adjectival clause and the principal clause are exchanged in the highest level.

The final reordered sentence is:

CS(S((സീത),V(സന്തോഷിച്ചു)),ADJC(CONDP(അപ്പോൾ),S(N(രാ
മൻ) , V(കൊന്നു), NG(NA( ), N(രാവണൻ))))

Reordered English sentence:

CS(S((Seetha),V(washappy)),ADJC(CONDP(when),S(N(Ra
man), V(killed), NG(NA( ), N(Ravanan))))

A depth first traversal of the reordered parse tree generates the following English sentence for the given input sequence:

Output : Seetha was happy when Raman killed Ravanan.

3.4 Cross Lingual Dictionary

The dictionary includes most of the commonly occurring verbs, nouns, pronouns, adjectives, inflectional and derivational suffixes, clause suffixes, etc. [9, 10]. Each entry in the file has three fields: the root word (morpheme), the morpheme tag and its translation. The verbs in past tense have their root words stored along with them. Since the system works with morphemes, the space required for the dictionary is less. The entries in the dictionary are shown in Table 2.

Table 2. Dictionary

Source word	Morpheme tag	Target word
പൂച്ച	Noun	cat
ഉടെ	Case suffix	's

Presently the system works for sentences which contains upto two adverbal or any number of adjectival clauses. The system can be modified to handle other sentences by adding appropriate grammar rules and transfer rules to the rule database. As the parser is a recursive one, it can handle sentences of any depth.

4. IMPLEMENTATION AND TESING

The system was implemented in Python language. The system was tested with a source file which contains more than 500 sentences. The sentences which follow the identified grammar rules in the rule database were correctly reordered. The system works correctly for simple sentences, complex sentences with two adverbal and any number of adjectival clauses and compound sentences.

5. SAMPLE OUTPUT

1. Malayalam sequence: "രമ", "പറഞ്ഞു", "അത്", "സീത", "കേട്ടു"

English sequence: "Rama", "told", "what", "seetha", "heard"

Reordered Malayalam sequence: CS(N (സീത) V(കേട്ടു) NC(NCE1(അത്),N(രമ),V(പറഞ്ഞു))

Output English morphemes: Seetha heard what Rama told

2. Malayalam sequence: "രാമൻ" , "രാവണൻ","എ", "ഉം", "മകൻ", "എ", "ഉം", "കൊന്നു"

English sequence: "Raman", "Ravanan", " ", "and", "son", " ", "and", "killed"

Reordered Malayalam sequence: S(N(രാമൻ) V(കൊന്നു)NC(NCC(N(രാവണൻ),NA(എ),CON(ഉം)) NCC(N(മകൻ),NA(എ),CON(ഉം))))

Output English sequence: Raman, killed, Ravanan, and, son, and

3. Malayalam sequence: "രാമൻ", "ഒരു", "നല്ല", "കുട", "സീത", "ക്ക്", "കൊടുത്തു", "അപ്പോൾ", "സീത", "സന്തോഷിച്ചു"

English sequence: "Raman", "a", "good" , "umbrella", "seetha", "to", "gave", "when" , "seetha", "was happy"

Reordered Malayalam sequence: CS(ADJC(CONDP(അപ്പോൾ) N(രാമൻ) V(കൊടുത്തു) NG(PA(ഒരു) PA(നല്ല) N(കുട)) NG(NA(ക്ക്)N(സീത)) N(സീത) V(സന്തോഷിച്ചു))

Output English sequence: When Raman gave a good umbrella to Seetha seetha was happy

4. Malayalam sequence: "രാമൻ", "സീത", "ക്ക്", "കൊടുത്ത", "കുട", "സീത", "കളഞ്ഞു"

English sequence: "Raman","seetha", "to", "gave", "umbrella", "seetha", "lost"

Reordered Malayalam sequence : N(സീത) V(കളഞ്ഞു) N(കുട) ADJC (N(രാമൻ) VRP(കൊടുത്ത) NG(NA(ക്ക്) N(സീത))

Output English sequence: Seetha lost umbrella Raman gave to Seetha

6. CONCLUSION

Translation methods like direct approach can be used only for translation between closely related languages like Hindi and Punjabi[11]. For translation between languages which belong to two language families like Dravidian and Indo Aryan, the syntactic structure transfer is a central design question. This paper discussed about the design and development of an efficient syntactic structure module for a Malayalam to English machine translator. The system uses a morphological parser for context sense disambiguation and chunking, syntactic structure transfer module and a bilingual dictionary. All the modules are morpheme based to reduce dictionary size. This system uses a rule-based architecture along with various linguistic knowledge components of both Malayalam and English. The set of rules for parsing and syntactic structure transfer are limited. The merit of the system is that the parsing and reordering algorithm is a general one and the rules are separate and can be added to the rule database without changing the parser. Additional modules like compound word splitter, finding and replacing collocations, finding and replacing named entities and handling of inter chunk and intra chunk dependencies can be added to this module to develop a machine translator from Malayalam to English [11, 12]. The results obtained are encouraging and the work can be extended to the creation of a full fledged machine translator from any Dravidian language to English since they all exhibit structural homogeneity.

7. REFERENCES

[1] Kumar G.B., Murthy K.N., UCSG Shallow Parser, Proceedings of CICLing-2006, Springer-Verlag, Volume 3878 , pp 156-167,

[2] Devi S.L., Ram V.S., et al., Syntactic structure transfer in a Tamil to Hindi MT system – A hybrid approach.

[3] Nair L R., Peter S.D. , Development of a Rule Based Learning System for Splitting Compound Words in Malayalam Language, IEEE Recent advances in Intelligent Computational systems (RAICS), 2011, pp. 751-755, 10.1109/raics.2011.

[4] Nair L.R, Peter S.D., Shallow Parser for Malayalam Language using Finite State Cascades , 4th International Congress on Image and signal processing, China, 2011, pp.2464-2467.

[5] Jurafsky D., Martin H..M., Speech and natural language processing, Prentice Hall, 2003, pp.657-690.

[6] Rich E., Knight K., Nair S.B, Artificial Intelligence, The McGraw Hill Companies , 2009, Third Edition pp. 295- 300.

[7] S. Abney, Partial parsing via finite state cascades, Journal of Natural Language Engineering, **2**(4), 1996, pp. 337-344.

[8] Abney S., Parsing Partial parsing via finite state cascades, Natural Language Engineering, 1995, Cambridge University Press, pp.1-8

[9] Idicula S.M, Peter S.D., A morphological processor for Malayalam language, South Asia Research. vol. 27 (2), pp.173-186.

[10] Pandian L., Geetha T.V, Morpheme based Language Model for Tamil Part of Speech Tagging, polibits, 38, pp.19-26, 2008, ISSN 1870-9044.

[11] Goyal V., Lehal G.S., Advances in Machine Translation Systems, Language in India www.languageinindia.com, 149 9 : 11, November 2009.

[12] Devi S.L., Pralayankar P, et.al., Verb Transfer in a Tamil to Hindi Machine Translation System, International Conference of Asian Language Processing, 2010, Harbin, China, pp.28-30.