Language Parsing and Syntax of Malayalam Language

Latha R Nair
School of Engineering
Cochin University of Science and Technology
latharnair@cusat.ac.in

David peter S
School of Engineering
Cochin University of Science and Technology
davidpeter@cusat.ac.in

***Abstract*— Parsers are integral components of many natural language processing systems for machine translation, language understanding etc. Parsers need the syntax of the language for creating the parse tree. This paper discusses the derivation of the syntax rules for sentences in Malayalam language. It also discusses the list of hierarchical syntax rules in context free grammar form. A set of part of speech tags and chunk tags were derived for representing the rules in context free grammar notation. The rule set covers the syntax of most of the commonly occurring sentences in Malayalam language.**

Keywords-parsing, Malayalam language, context free grammar, syntax etc.

I. INTRODUCTION

The process of generating the sentence through derivation using a set of grammar rules is called parsing and the generated hierarchical structure is called the parse tree of the sentence. The parser for a language needs the syntactic structure of the sentences of the language. The part of speech(POS) tag set for various words in the sentence, the groups of co-occurring words known as word chunks, the structure of sentences in a language and the hierarchical dependencies of chunks in sentences are required for the derivation of the syntax of sentences[1].

II. PREVIOUS WORKS

Context free grammar based has been used for top-down parsing of Myanmar sentences [2]. A probabilistic method has been tried for parsing natural language sentences [3,4]. A top-down parsing algorithm to accommodate ambiguity and left recursion in polynomial time has also been tried [5]. A shift reduce parsing technique has been used for word sense disambiguation [6].

III. LANGUAGE CHARACTERISTICS

In order to arrive at a computational grammar for the language the set of word classes (Part Of Speech tagset), chunk tagset and the hierarchical dependencies among the chunks are needed. This requires a careful analysis of the different classes of sentences in the language.

Both morphology and morphotactics of the language have been considered for this purpose. Malayalam is a highly agglutinative language and the morphological variations are more for the language compared to English or Hindi. The nouns have inflections due to case, gender and number information. The verbs are inflected due to tense, aspect and mood information. In Malayalam language the following set of sentence classes are found. i)simple sentence ii)complex sentence and iii)compound sentences. The sentences may contain clauses. The clauses found in the language are i) adjective clause ii)adverb clause and iii) noun clause.

IV. SELECTION OF POS TAGS

First step in deriving the syntactic structure of Malayalam sentences was the identification of set of word categories in a Malayalam sentence called part of speech tags. Lexicalized tags are very useful for machine translation systems and language understanding systems [7,8]. Since we found that a morpheme based parsing was appropriate for a highly agglutinative language like Malayalam it was decided to give a unique tag name for each morpheme category. The inflectional and derivational suffixes were given separate tag names. The set of tags identified for our problem are listed in Table 1.

V. SELECTION OF CHUNK TAGS

After selection of POS tags in sentences the chunk tags were identified. The syntax rules are to be used by a parser for a lexicalized tree adjoining grammar (LTAG) based machine translation system from Malayalam to English language. So the chunks that are to be rearranged for the translation from Malayalam to English were identified and given a unique tag name for each chunk. The tagset includes all of the tags in IIIT tagset and also some additional tags to handle higher level constructs like clauses and sentences. The list of chunk tags identified is shown in Table 2. A chunk tag is allotted for each of the morpheme group found in the hierarchical structure for the sentences in Malayalam. The tags were so chosen that it forms the morpheme groups to be used in the reordering process to generate the target language parse tree during the translation process[9,10].

TABLE I POS TAGS

No.	Tag	Description
1	PL	Plural suffix
3	NA	Postposition
4	PA	Adjective
5	N	Noun
6	V	Verb
7	ADJA	Adjectival suffix
8	ADVA	Adverbial suffix
9	PAV	Adverb
10	VN	Verbal Noun

11	V RP	Relative participle suffix
12	NCA	Noun clause suffix
13	ADVCA	Adverbal clause suffix
14	INFA	Infinitive suffix
15	DJ	Disjunction
16	C	Conjunction
17	LOC	Locatives
18	VA	Verbal suffix

VI. Hierarchical Dependancy Structures

Clauses in a sentence can be nested one inside the other, resulting in a hierarchical or tree like structure. This aspect of structure is called the hierarchical structure [11,12]. Clauses in a sentence are not completely independent of one another but there are inter-clause dependencies. For example, a noun phrase being modified by a relative clause has two roles to play, one in the relative clause and the other in the outer clause.

According to Universal clause structure grammar (UCSG) all inter-clause dependencies systematically flow down the clause structure tree from the root towards the leaves [13,14]. Also, the constituents of a clause do not cross clause boundaries in scrambling. Verb groups and sentinels

Table II Chunk Tags

No.	Tag	Description
1	NP	Noun Group
2	VG	Verb Group
3	NC1	Noun clause
4	ADVC	Adverb clause
5	ADJC	Adjective clause
6	NPC	Conjunct Noun
7	S	Sentence
8	CS	Compound sentence
9	CMPN	Compound noun
10	ADJCNP	Adjectival clause + Noun
11	ADJG	Adjective group
12	INFSG	Infinitive + verb group
13	INF	Infinitive
14	ADVG	Adverb group
15	VGC	Compound verb
16	VA	Verbal suffix
17	ADJLOC	Locative adjective

contain all the required information for recognizing clauses, for determining the nested or hierarchical structure of clauses and for determining the clause boundaries. It is seen that every clause in a sentence except for the main clause has a sentinel which marks one of the boundaries of that clause. The sentinel marks either the beginning or the end of the clause depending upon the language in use. Also every clause must have exactly one verb group.

Malayalam belongs to Indo- Dravidian family of languages and it is a relatively free word order language like other Dravidian languages. Malayalam is 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 freedom in the order of constituents. Normally the verb remains in the sentence final position. Word order is less important mainly because noun groups are marked for cases and the verb agrees with the subject in gender, number and person. Subjects and objects are often dropped. The subject of a sentence is expressed by a noun group in the nominative case in most of the sentences. Normally all modifiers precede the modified [15].

There are a variety of subordinate clauses. Subordinate clauses also precede the main clause. They are normally non-finite forms of verbs which occur in the clause final position and mark the right hand boundary of the respective clauses. All these assertions were used to form the syntax 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.

VII. Hierarchical Dependancy Rules For Chunks In Malayalam Language

The set of Hierarchical dependency rules for chunks in Malayalam language identified are given in Table 3. The rules are given in context free grammar form. Rules for forming chunks are given below with examples. A transliteration of Malayalam sentence and its English translation are given.

1) ***Start - Highest level chunk***

1. S - A simple sentence
2. CS – Complex sentence

2) ***CS - Complex sentence***

1. An adverb clause followed by a simple sentence

T: (r*aamu padichaal*) (ADVC) *(pareekshayil vijayikkum*) (S)

E: If Ramu studies he will pass in the examination

2. A noun clause followed by a complex sentence

T: (r*aaman mOhane adichchennu)*(NC) (*ramaye kandappOL seetha paRanjnju*)(CS)

E: When Seetha saw Rama she told that Raman hit Mohan 3.An adverb clause followed by a complex sentence

4. A noun clause followed by a simple sentence

3) S - Simple sentence

One or more noun groups followed by a verb group.

E:(Raman hit Mohan)

T:NP*(raaman*) NP(*mohane)* VG(*atichchu*)

4) ADVC - Adverb clause

A simple sentence followed by adverb clause marker.

T: (S(*raamu vann*) CONDP(*aal))*

E: If *Ramu* comes

5) NC1 - Noun clause

A sentence followed by the clause marker *ennz* forms noun clause.

T: ((*rama vannu*)(S) *ennu*(NCE1) (*mOhan paRanjnju*)(S))

E: (Mohan told that Rama had come)

TABLE III . HIERARCHICAL DEPENDENCY RULES

Sl. No	Production rules
1	START=>S|CS
2	CS=>ADVC S|NC1 S
3	S=>NP^{+} VG
4	ADVC=>S ADVCA
5	*NC1=>S NCE1*
6	NPC1=>NP C NPC=>NPC1 NPC1|NPC1 NPC1 NPC1*
7	ADJC=>NP* VRP
8	NP=>ADJG* N|ADJG* N NA|ADJG* N PL NA|ADJG* N PL|ADJG* NPC|ADJG* NC2 NA|ADJC NP|ADJLOCN ADJLOCN=>ADJLOC N
9	CMPN=>N N
10	ADJCNP=>ADJC NP
11	ADJG=>PA|N ADJA | ADJLOCADJLOC=>N LOC
12	VG=>ADVG* V NE|ADVG* VG1|ADVG*V|INFSG|INFG|ADVG* V QA| N CVA
13	INFSG=>INF V | INF V VA
14	INF=>V INFA
15	*ADVG=>PAV|N ADVA*

6) ***NPC - Noun Conjunct***

A noun group followed by the conjunct suffix *um* forms a conjunct noun.

rama(NP) – *um*(C) ravi (NP)– *um* (C) (Rama and Ravi)

7) ***ADJC - Adjective clause***

A sentence followed by relative participle forms an adjective clause.

T: ((*seetha paRanjnja*)(ADJC) *kadha Ramakku ishtappettu*)S

E: (Rama liked the story which Seetha told)

8) ***NP - Noun chunk***

1.A noun alone.

(T: *raaman* / E: Raman)

2.A noun followed by a case marker

(T: *raaman-Odu* / E: to Raman)

3.A noun followed by a plural marker and a case suffix

(T :*kutti-kaL-Odu* / E: to children)

4.A noun preceded by an adjectival clause

T: (rama paRanjnja)(ADJC) kaTha(N)

E: (the story which Raman told)

9) CMPN - Compound noun

A noun followed by another noun.

(T: *vivaaha-mOthiram* / E: wedding ring)

10) ADJCNP - Noun preceded by an adjective clause

The adjective clause and the noun it qualifies are grouped as they are to be treated as a single unit during structure transfer from Malayalam to English.

11) ADJG - Adjective chunk

1.A pure adjective

(T:*nalla* / E: good), (T:*kure* / E:some)

2.A derived adjective formed by a noun followed by adjectival suffixes.

(T: *bhangi / E:* beautiful) – *(ulla*)(Adjectival suffix)

12) VG - verb group

1. Zero or more adverb group followed by a verb, verb and inflectional suffixes or verb, inflectional suffix and question tag.

(T: *pOyi*/ E: went)(V), (T: *pOk*)(V) – (*unnu* /is going)(VA)

2. A Compound verb i.e. a verb followed by another verb

chaadi (V) *kayar*i(V) (climbed jumping), *Odi*(V) *pOyi*(V)(went running)

3. Infinitive followed by a verb

pOk(V)-*aan*-(INFA) *pOyi*(V) (went to go)

13) INFSG - Infinitive followed by a verb group

The infinitive and the verb following it are grouped.

pOkaan(INF) *thutangi*(V)(started to go), *vaangaan*(INF) *pOy*i(V)(went to by)

14) INF- Infinitive

A verb followed by the suffix *aan* is taken as infinitive.

pOk(V) – *aan*(INFA), *var*(V)- *aan*(INFA)

15) ADVG - Adverb group

1. Pure adverb (PAV)

*pathukk*e(slowly), *pettennu*(quickly)

2. Noun followed by adverbial suffix

bhangi(N)- *aayi*(ADVA)(beautifully)

16) VGC- Compound verb

A verb followed by another verb are grouped to form a compound verb.

chaati(V) – *kayaRi*(V), *natannu*(V) – *pOyi(V)*

VIII. CONCLUSION

The paper discussed the derivation of the syntactic structure of sentences in Malayalam language. The set of POS tags, chunk tags and the set of hierarchical dependency rules identified cover most of the commonly occurring sentence classes in Malayalam. The rule set can be used by the parser module for a machine translation system from Malayalam to any other language like English with wide syntactic structure difference.

REFERENCES

[1] Aravind K. Joshi, L. Levy and M. Takahashi,Tree Adjunct Grammars, Journal of Computer and System Sciences, volume10, issue1, p.p.136-163, 1975.

[2] Win Win Thant, Tin Myat Htwe et. al., Context Free Grammar Based Top-Down Parsing of Myanmar Sentences, International conference on computer science and information technology, Pattaya, p.p. 71-75, 2011.

[3] Mark A Jones et. al., A Probabilistic parser applied to software testing documents, Proceedings of national conference on Artificial Intelligence, San Jose, p.p. 322-328, 1992.

[4] Brian Roark, Probabilistic top down parsing and language modeling, Computational linguistics, volume 27, p.p. 249-276, 2001.

[5] Richard A. Frost, Rahmatullah Hafiz, A new top-down parsing algorithm to accommodate ambiguity and left recursion in polynomial time, ACM SIGPLAN, volume41, issue5, p.p. 46-54, 2006.

[6] Stuart M Scheiber, Sentence disambiguation by a shift reduce parsing technique, 8th international Joint conference on artificial intelligence, p.p. 699-703, West Germany, 1983.

[7] A.Abeille, et. al., Using lexicalized tags for machine translation, 13th International conference on computational linguistics, volume 3, Finland, p.p. 1-6, 1990.

[8] Murthy. K. 2002. MAT: A Machine Assisted Translation System. In Proceedings of Symposium on Translation Support Systems, STRANS-2002, IIT Kanpur, India,. p.p. 134-139, 2002.

[9] Stuart M Shieber, Yves Schabes, Generation and synchronous tree adjoining grammars, Computational intelligence, 1992, p.p. 220-228.

[10] Steve Deneefe, Kevin Knight, Synchronous tree adjoining machine translation, EMNLP-2009: Proceedings of the 2009 Conference on Empirical methods in natural language processing, Singapore, p.p. 727-736, 2009.

[11] Noam Chomsky, On Certain Formal Properties of Grammars, Information and Control, Vol. 9, p.p.137-167, 1959.

[12] Noam Chomsky, Syntactic structures, 2nd edition, ISBN_3_11_0 17279_8, 1957.

[13] K. Narayana Murthy, A. Sivasankara Reddy, Universal Clause Structure Grammar, Computer Science and Informatics, Vol. 27, No 1, Special Issue on Natural Language Processing and Machine Learning, p.p. 26-38, 1997.

[14] Murthy K.N, UCSG and the syntax of relatively free word order languages, South Asian Language Review VII, 1997

[15] E.V.N.Namboothiri, VakyaGhatana, Kerala bhasha institute, third edition, 1997

.