VASISTH-An Anaphora Resolution System

Sobha,L
M.G.University
Kottayam, Kerala, India
sobha@rocketmail.com

B.N.Patnaik
Indian Institute of Technology
Kanpur,UP, India
patnaik@iitk.ac.in

Introduction

Most of the anaphora resolution systems developed so far are monolingual, ie for particular languages and not easily extendable to other languages. VASISTH, in contrast, is a multilingual system, which presently handles two different languages from two different language families: Malayalam, from, Indo-Dravidian and Hindi from Indo-Aryan. It can easily be extended to handle other Indian languages, more generally, other morphologically rich languages. What distinguishes VASISTH from other similar systems is that exploring the morphological richness of the Indian languages, it makes limited use of syntax and uses only morphological markings to identify subject, object, clause etc. It uses limited parsing: the information required from the parser is limited to parts of speech tagging, clause identification, subject of the clauses and person-number-gender of the NPs. Initially VASISTH was developed and tested for Malayalam, and then modified for Hindi. It is sensitive to ambiguity occurring in pronoun resolution but does not resolve the ambiguity. The system can resolve all referentially dependent elements. This system resolves all referentially dependent elements such as pronominals, non-pronominals, gaps and ellipsis. Pronominals are classified into two: pronouns and one-pronouns and non-pronominals are classified into three: reflexives, reciprocals and disributives. Gaps are of two types: forward and backward. This paper confines itself to the handling of only one-pronouns, distributives, reciprocals and gaps in both the languages.

Anaphors in Malayalam and Hindi

The third person Pronouns in Malayalam (M) and Hindi (H) are as follows:

Singular	Plural

avan(he)
aval(she)
atu(it)
	avar(they)

Singular	Plural
voh ve(honorific)	ve

In Malayalam there is a singular-plural distinction in all third person pronouns and a masculine-feminine distinction where as in Hindi there is no distinction. Both in Malayalam and Hindi the pronouns take the entire range of cases. The relationship of pronouns with their antecedents in Malayalam and Hindi are described below.

1.(M) mo:han$_i$ avanRe$_i$ kuttiye kantu.
mohan he-poss child-acc see-pst
(Mohan saw his child.)
2.(M) mo:han avane$_i$ aticcu ennu kRisnan$_i$ paRannu.
mohan he-acc beat-pst compl krishnan say-pst
(Krishnan said that Mohan beat him.)
3.(M)*mo:han$_i$ avalute$_i$ ammaye kantu.
mohan she-poss mother-acc see-pst
(Mohan saw her mother.)
4.(H) ra:m$_i$ uski$_i$ kita:b sya:m ko di.
ram he-poss book syam-acc give-pst
(Ram gave his book to Syam.)
5.(H) ra:m$_i$ ko ma:lum hai ki mo:han uskeliye$_i$ ka:m nahi kare ga.
ram–acc know-prst copula compl mohan he-ben work neg do-prst
(Ram said knows that Mohan does not work for him.)

In (1) pronoun is in the possessive form and its antecedent is “mo:han” “Mohan” which is also the subject of the sentence. In (2) the antecedent of the accusative pronoun “avane” “him” is “kRishnan” “Krishnan” which is the subject of the immediate clause IC (immediate clause is any clause which precedes or follows the clause which contains the anaphor)in which the pronoun occurs. In sentence (3) the sentence is ungrammatical in the given reading because there is a gender disagreement between the pronoun “avalute” “her” which is feminine and the antecedent “mo:han” “Mohan” which is masculine. From this we can arrive at the following:

I. A pronoun P is coreferential with an NP iff the following conditions hold:

a. P and NP have compatible P, N, G features.
b. P does not precede NP.
c. If P is possessive, then NP is the subject of the clause which contains P.
d. If P is non-possessive, then NP is the subject of the immediate clause which does not contain P.

In Hindi sentence (4), the antecedent of the possessive pronoun “uski” is the subject of the sentence “ra:m”, and in (5), the non possessive pronoun is “uskeliye” “his” and the antecedent is “ra:m” “Ram” which is the subject of the IC. From this we can arrive at the following

II. A pronoun P is coreferential with an NP iff the following hold:

a. P does not precede the NP.

b. If P is non-possessive, then NP is the subject of the IC in which P does not occur.

c. If P is possessive, then NP is the subject of the IC in multiple clause constructions or the NP immediately preceding the P.

The one-pronouns in Malayalam and Hindi can be classified on the basis of countability, which is an inherent feature, the one-pronoun in Hindi are of two types: The pronoun, which are [+C], and those which are [-C]. The two instances of one-pronoun, which are homophones have the form “kuch” [+C], “kuch” [-C]. The [+C] pronoun takes [+/-animate, +count] NPs as its antecedent and [-C] pronoun takes [-animate, -count] NPs. In the case of homophone pronouns the antecedent is [+/-animate, +/-count] NPs. The antecedent NP is the non-subject NP in the immediate clause in which one-pronoun does not occur. Here we have to use world knowledge, which is outside the purview of this work. These are featured in the tabular form below.

One-Pronoun in Malayalam and Hindi	**Inherent feature of One-Pronoun**	**NP which can be the antecedent**
“orannam” “ ek” “one”	+C	+/-animate, +count
“alpam” “thoda” “little”	-C	-animate, +count
“kuRe” “kuch” “some”	+C, -C	+/-animate, +count

Given below are examples which demonstrate the above claim regarding [+C] pronouns along with two other claims that hold for [-C] pronouns, namely that (a) the antecedent must precede the pronoun and (b) the one-pronoun must have an explicit antecedent.

6.(M) na:n rantu paksikale$_i$ ku:ttil kantu. ra:man orannatte$_i$ marattil kantu.
I two bird-pl-acc nest-loc see-pst raman one tree-loc see-pst
(I saw two birds in the nest. Raman saw one on the tree.)

7.(M) ra:man pa:lu$_i$ vanniccu pu:cca alpam$_i$ kuticcu.
raman milk buy-pst cat little drink-pst
(Raman bought milk. The cat drank a little.)

8.(M) kRisnan ra:manu a:pple$_i$ koduttu avan kuRe$_i$ kalyiccu.
krishnan raman-acc apple give-pst he some eat-pst
(Krishnan gave Raman an apple. He ate some (part of it).)

9.(M) ammama:r kuttikalkku vellam$_i$ koduttu avar kuRe$_i$ kuticcu.
mother-pl children-pl water give-pst they some drink-pst
(Mothers gave water to the children. They drank some.)

10.(H) mai ne do cidiya$_i$ ghosle me dekhi, mo:han ne ek$_i$ sakha me dekhi.
I-erg two birds nest-loc see-pst mohan-erg one branch-loc see-pst
(I saw two birds in the nest, Mohan saw two on the branch.)

11.(H) mo:han du:dh$_i$ kharida, billi tho:da$_i$ pi:liya.
mohan milk buy-pst cat little drink-pst
(Mohan bought milk, the cat drank a little.)

12.(H) aurate:m bacco:n ko pa:ni$_i$ diya ve kuch$_i$ pi:ye.
woman-pl children- acc water give-pst they some drink-pst
(Women gave some water to the children, they drank some.)

13.(H) mo:han ne ra:m ko se:v$_i$ diya ve kuch$_i$ kha:ya.
mohan-erg ram-acc apple give-pst he some eat-pst
(Mohan gave Ram some apple they ate some.)

In (6) "orannatte" “one” refers to the birds in the previous sentence and in (17) “orannum” refers to the books. The one-pronoun “alpam” in (7) refers to “pa:lu” “milk”. In sentence (6) the one-pronoun is [+C] and in (7) it is [-C]. In (8) the one-pronoun “kure” “some’ refers to “a:pple” “apple” which is [+C] and (9) it refers to “vellam” “water” which is [-C]. The sentences in Hindi also show the distribution of one-pronouns. In (10) “ek” “one” refers to the “birds” in the previous sentence. In (11) the One-pronoun “tho:da” “little” refers to “du:dh” “milk”In (12) the one-pronoun “kuch” “some” refers to “pa:ni” “water”, which is [-C] and in (13) “kuch” “some” refers to apple which is [+C]. From the above we arrive at the following:

II. A one-pronoun corefers with an NP iff

e. Non-subject NP is in IC of one-pronoun.
f. NP precedes the one-pronoun.
g. one-pronoun and NP agree with respect to C features.

Consider the non-pronominals in both the languages. There are three types of non-pronominals in these languages, emphatic, non-emphatic and the possessive. The non-emphatic reflexive in Malayalam is “ta:n”, which can take all the case forms. In Hindi there are four non-emphatic reflexives and they are “apna”, “apnea:p”, “khud”, “svayam

". The following examples show the non-emphatic reflexive-antecedent relation in both Malayalam and Hindi

14.(M) ra:man$_i$ paRannu tanne$_i$ na:ttuka:r aticcu ennu.
raman say-pst self-acc people hit-pst compl
(Raman said that people hit him.)
15.(H) mo:han$_i$ ne apnea:p$_i$ ghar cala gaya.
mohan-erg refl house go-pst
(Mohan went home by himself.)
16.(H) mo:han ne apne ko aine me dekha
mohan-erg refl–acc mirror-loc see-pst
(Mohan saw himself in the mirror.)
17.(H) mai apni ladki ki intazar me hum
I refl girl-acc wait–loc copula
(I am waiting for my daughter.)
18.(H) manisa ne apnekeliye ghar kharida
manisha–erg refl-ben house buy-pst
(Manisha bought a house for herself.)

In sentence (14) the non-emphatic reflexive "tanne", which is in the accusative form, has its antecedent "ra:man" "Raman" in the matrix clause. In the next two sentences (15) and (16) too, the antecedent is the subject of the matrix clause. In (16) the antecedent of "apne ko" is the subject of the clause "mo:han" "Mohan". The same is the case with (17), where antecedent of "apni" is "mai". Sentence (18) has "manisa" "Manisha" as the antecedent for the non–emphatic reflexive "apnekeliye", which is the subject of the clause. From the above examples we arrive at the following.

III. A non-emphatic reflexive R$_2$ corefers with an NP iff the following holds:
h. NP is the subject of the clause in which the R$_2$ occurs.
k. With an N if the NP is possessive then the head noun is the antecedent of R $_2$

Turning to the other type of reflexive, the emphatic reflexive in both the languages. In Malayalam the emphatic reflexives are "tannata:n" and "svayam", the latter borrowed from Sanskrit and it is morphologically invariant in that it does not take any case forms. In the case of "tannata:n" it takes only two cases namely, nominative and accusative. Both "svayam" and "tannata:n" have free distribution. The emphatic reflexives in Hindi are "apnea:p", "kudh" and "svayam". The following examples show the occurrence of emphatic reflexives in both the languages.

19.(M) ra:man$_i$ tannata:n$_i$ sku:lil po:yi.
raman self school-loc go-pst
(Raman went to the school himself.)
20.(M) ra:man$_i$ svayam$_i$ sku:lil po:yi.
raman self school-loc go-pst
(Raman went to the school himself.)

21.(H) bacce$_i$ ne apnea:p$_i$ khana khaya.
children–erg refl food eat-pst
(Children ate their food by themselves.)
22.(H) manisha$_i$ khud$_i$ a:yi thi.
manisha refl come-pst
(Manisha came by herself.)
23.(H) bhagava:n$_i$ svayam$_i$ prakat ho gaye.
God refl manifest pst
(God himself became manifested.)

In the above sentences (19) and (20), the antecedent of the emphatic reflexives “tannata:n” is the subject of the clause in which it occurs, that is “ra:man” “Raman”. The same is the case with the emphatic reflexive “svayam”. Both “svayam” and “tannata:n” have free distribution. In (21) the antecedent of the emphatic reflexive apnea:p is “bacce” “child” and it is the subject of the clause in which it occurs. Same is applicable for (22) where “manisha” “Manisha” is the antecedent of “khud”. In (23) the antecedent of “svayam” is “bhagava:n” which is the subject of the clause in which it occurs. Thus it follows:

III An emphatic reflexive R$_1$ corefers with an NP iff
i. NP is the subject of the clause in which R$_1$ occurs.
Corefers with an N, the head of possessive NP,
j. If the NP is the subject of the clause that contains R$_1$.

The third type of reflexive is the possessive reflexive, which is found only in Malayalam is “svantam” which is also borrowed from Sanskrit. It behaves like the emphatic reflexive “tannata:n”. Consider the following examples that contain the possessive reflexive “svantam”:

24.(M) si:ta$_i$ svantam$_i$ kuttiye aticcu ennu amma paRannu.
sita self child-acc beat-pst compl mother say-pst
(Mother said that Sita beat her (=Sita's) child.)
25.(M) ra:man$_i$ svantam$_i$ vanti kRisnanu o:tikka:n kotuttu.
raman self vehicle krishnan-dat drive-purposive give-pst
(Raman gave his (=Raman's) vehicle to Krishnan for driving.)

In both the cases the antecedent of the possessive reflexive “svantam” is the subject of the clause in which “svantam” occurs. In (24) “si:ta” “Sita” is the antecedent of “svantam” and in (25) it is “ra:man” “Raman”. From the above we conclude that:

V. The possessive reflexive R$_2$ corefers with an NP iff
k. NP is the subject of the clause which contains R$_2$.

Now consider the reciprocals and distributives in both the languages. There are several types of reciprocal anaphors in Malayalam. The ones which have the most frequent distribution are “ora:l-ora:l” and “ora:l-matte-a:l”. There are other reciprocals in Malayalam and they are “anyo:nyam”, “tammiltammil", “parasparam” and “anno:ttum-inno:ttum”. These forms do not take any case markers. The reciprocals in Hindi are “paraspar” and “ek dusre”. The following examples show the relationship between the antecedent and the reciprocal.

26.(M) ii kuttikal$_i$ ora:l-ora:le/ora:l-matte-a:le$_i$ atikkilla.
these children each other beat-neg
(These children do not beat each other.)
27.(M) avarkku$_i$ ora:lkku-ora:le/ora:lkku-matte-a:le$_i$ istamilla.
they-dat each other like-neg
(They do not like each other.)
28.(M) avar$_i$ anyo:nyam/ tammil tammil/ parasparam/ anno:ttum inno:ttum$_i$
they each other
sne:hiccu.
like-pst
(They liked each other.)
29.(H) ye bacce$_i$ ek dusare$_i$ se ba:t nahi karenge.
These children eachother-acc talk neg do-fut-pl
(These children will not talk to eachother.)
30.(H) ye$_I$ ek dusere$_I$ ko pasand nahim karte hai.
These eachother-acc like neg do-prst-pl copula
(They do not like each other.)
31.(H) un$_i$ dono me paraspar$_i$ ladhai hua.
This two-loc each other war be-pst
(There was war between them.)

The antecedents of the reciprocal anaphor in (26) and (27) are “kuttikal” “children” and “avarkku” “they” respectively. In (26) and (27) each the antecedent precedes the anaphors. The sentence (28) shows the distribution of other reciprocals in Malayalam. In sentence (29) the antecedent of “ek dusre” is “bacce” “children” which is plural and precedes the reciprocals. In sentence (30) the antecedent is “ye” “these” which again is plural and precedes the reciprocal. Unlike Malayalam, reciprocals in Hindi take all case suffixes. Consider the other reciprocal “paraspar”. In (31) the antecedent of the reciprocal “paraspar” is “un dono”. Here also the antecedent is plural and precedes the reciprocal. The reciprocal “paraspar” does not take any case suffix. The coindexing in the above examples shows that the antecedent of the reciprocal anaphor is the subject of the clause in which it occurs. The antecedent has to be plural for a reciprocal anaphor. It is also evident that the antecedent must precede the anaphor.

VI. A reciprocal anaphor R’ is said to corefer with an NP iff
k. NP is the subject of the clause, which contains R’.
l. NP is plural.
m. NP precedes R’.

Now consider the distributive anaphors. In Malayalam distributives are “avar-avar” and “avan-avan”. These distributive anaphors are reduplication of the pronouns “avar” and “avan” and with respect to antecedents, they behave like reciprocals. The distributive reflexive “apna apna” is the reduplicated form of the non-emphatic reflexive “apna” and behaves like reciprocal.

32.(M) ammama:r$_i$ avar-avarute$_i$ kuttikale raksiccu.
mother-pl their children save-pst
(Mothers saved their own children.)
33.(M) ellavarum$_i$ avan-avanRe$_i$ saukaryam ma:tram no:kki.
all their convenience alone see-pst
(All looked at their own convenience.)
34.(H) mo:han aur ra:m$_i$ ko apne-apne$_i$ ghar pasand hai.
mohan and ram-acc each other house like copula
(Mohan and Ram like their respective houses.)
35.(H) mo;han aur ra:m$_i$ ne si:ta aur gi:ta ko apni-apni$_i$ kita:b di.
mohan and ram-erg sita and geeta-acc each other book give-pst
(Mohan and Ram gave their respective books to Sita and Geeta.)

In (32) and (33) the antecedent of the distributive anaphor is the subject of the clause in which it occurs. In (32) the antecedent is “ammama:r” “mothers” and in (33) it is “ella:varum” “all”. In both the cases the antecedent is plural and it precedes the anaphor. In sentence (34) the antecedent of the distributive reflexive is “ra:m aur mo:han” “Ram and Mohan”, which is the subject of the sentence and is plural. It precedes the distributive reflexive. In sentence (35) the antecedent is “ra:m aur mo:han” “Ram and Mohan” which is plural, which precede the distributive reflexive and also is the subject of the clause in which the distributive reflexive occurs. The distributive reflexive does not take any case suffixes in Hindi. It takes the gender of the subject NP in a non-ergative sentence and that of the object NP in an ergative sentence. In conclusion it can be stated that

VI. A distributive reflexive D is coreferential with an NP iff the following holds:

o. NP is plural.
p. NP precedes D.
q. NP is the subject of the clause in which D occurs.

Now we consider the gaps in Malayalam and Hindi. They occur in discourses, each with at least two parallel structures elided for the reasons of economy or emphasis and can be understood in terms of the corresponding constituent in the other of the parallel constructions. Gaps are different from ellipsis in the following ways:

The gaps have the following properties:

1. A constituent such as subject, object or verb is omitted to avoid repetition.
2. Occur in intra-sentential constructions.

The gaps are of two types: forward and backward. In forward gapping, the gap occurs in the initial clause whereas in backward gapping, the gap occurs in the second clause. The gapped entity can be the subject NP, the object NP, the verb or the whole VP. Consider the following examples, which have forward gapping:

36.(M) sya:m kuttikale sne:hikkunnu pakse avanRe bha:rya verukkunnu.
syam children like-prst but he-poss wife hate-prst
(Syam likes children but his wife hates.)
37.(H) si:ta ne ro:ti khayi, ca:y pili.
sita-erg roti eat-pst tea drink-pst
(Sita ate roti, drank tea= Sita ate roti and drank tea.)

The above sentences contains two parallel constructions. The parallel constructions are identified by the presence of the coordination markers. Here the coordinate marker is "pakse" "but" and "aur" "and". In (36) the element gapped is "kuttikale" "children" which is the direct object and in (37) the element gapped is "khayii" "eat" which is a complex verb of the coordinate marker.
Now consider the following examples which depict backward gapping:

38.(M) ra:man si:tayeyum hari ritayeyum kalya:nam kaliccu.
raman sita-acc-coord hari rita-acc-coord marry do-pst
(Raman married Sita and Hari married Rita.)
39.(H) mo:han ne si:ta se aur hari ne gita se sa:di ki.
mohan erg sita-comm and hari-erg gita- inst marry-pst
(Mohan married Sita and Hari married Gita.)

In (38) the complex verb "kalya:nam kaliccu" "married" is gapped and In (39) the element gapped is "sa:di ki" "married" which is a complex verb
From the above examples we can arrive at the following regarding forward and backward gapping:

I. If Q is a sentence and Q' and Q'' are the two parallel structures which constitute Q, then the verb V or the noun phrase NP occurs recursively iff
   a. for forward gapping
      q. NP is any constituent in Q'.
      r. V is in Q'.

   b. for backward gapping
      s. NP is the subject of Q''.
      t. V is in Q''.

Coming to ellipsis next, we consider inter-sentential ellipsis involving wh constructions (wh-const) and also question constructions (q-const) where the "o:", the question morpheme occurs at the end of a declarative sentence. The elided material can be any constituent and unlike gaps even non-constituents can be elided. In the case of wh constructions the material that cannot be elided is the OBJECT (for the present purpose, OBJECT is used to refer to what is called "discourse focus"). Consider the following examples:

40.(M) ni: evite   po:yi?
          you   where   go-pst
          (Where did you go?)

          vi:ttil.
          house-loc
          (To the house. (=I went to the house.))

In the above sentence the wh word is "evite" "where" which is locative. The response sentence to this wh-construction has two constituents elided, "na:n" "I" and "po:yi" "went". The former is the subject and the latter is the verb. The constituent that is not elided in the response is the locative, which is the focus of the sentence. The following table gives the information about wh-word and its focus:

Wh-word	**Focus**
etra	Nominative
a:ru	Nominative
entine	Accusative
entukontu	Instrumental
enno:ttu	Dative
ennane	Locative
eppo:l	Locative
evite	Locative

Another type of ellipsis that Malayalam has is yes/no question constructions as in the following:

41.(M) ni: kalicco: ?
          you   eat-pst-Qmorph
          (You ate? (=Did you eat?))

          illa.
          no
          (No. (=I did not eat.))

Here the question is formed by adding a question morph to the verb or to the noun. The response will be either "illa" "no", "a:nu" "yes", "uvvu" "yes" or the verb without the question morph. Consider the following example:

42.(M) kalico:?
eat-que morph
(Ate? (=Did you eat the food?))

uvvu.
Yes
(Yes. (=Yes, I ate the food.))

or

kaliccu.
eat-pst
(Ate. (=Yes, I ate the food.))

or

uvvu, kaliccu.
yes eat-pst
(Yes, Ate. (=Yes, I ate the food.))

There are three possible responses, in each of which different material is elided. In the first response only one constituent is present and it is "uvvu" "yes". In the second the verb "kaliccu" "ate" is the only constituent present and in the third response both "yes" and the verb are present.

The question (41) itself has elided material. If the clause is a wh construction then the subject is elided and if a q-construction, then the subject, the object or both. From the above we arrive at the following:
If the clause is a wh-construction, then the elided fragment in the response is of the following:

u. Subject.
v. Verb.
w. Both the subject and the verb.

X. If the clause is a q-const then, the elided fragment in the response can be one of the following:

x. Subject.
y. Object.
z. Both the subject and the object.

Coming to ellipsis in Hindi. Unlike Malayalam, Hindi does not have q-const constructions. Consider the following examples.

43. tum kaha gayi thi?
you where go-pst
(Where did you go?)

sku:l
school
To school (I went to the school.)

In (43) there are two constituents elided, "tum" "you" and "gayi thi" "went". The former is the subject and the latter is the verb. The constituent that is not elided in the response is the objective/locative, which is the focus of the sentence and the wh word "kaha" "where" takes a nominative focus. The following table gives the wh word and the possible focus it can take.

Wh-word	Focus
kitna	Nominative
kisse, kon	Nominative
kab	Nominative
kya, kis ko	Accusative
kyo:m	Verb
kaha:m	Locative
kis se	Locative

VIII. If S is a wh-construction, then the elided fragment in the responds can be of the following:

w. the subject
x. the verb
y. both the subject and verb.

To conclude, we have dealt with different types of anaphors and ellipses in Malayalam and Hindi.

The Parser.

The parser described here is rule-based, not principle based, and in any case does not use any of the contemporary models of grammar, such as GB, Minimalist Grammar, LFG, etc. We start with the hypothesis that for configurational languages like Malayalam and Hindi, rule-based parsing is easier and faster than principle-based parsing. The parser does not yield tree structures as output; the output is sequential (linear) which show how a sentence is broken into subparts and how the subparts are broken up into smaller parts in turn. Keeping in view the requirement of simplicity for parsing, the grammar (as explicated in the earlier chapters) has been slightly modified occasionally. For example, although copula is really unrelated to the "dativeness" of the subject, it has been used as a subject identifier, for the purpose of parsing. Similarly, possessive NP has been used as a subject-identifier, although the grammatically more correct statement would not characterize "possessive NP" as a subject identifier. Again, all NPs, other than the subject are classified, for parsing purposes, as object NPs. It is a well-known fact that the complementizer occurs in the (embedded) clause final position, however, the rules of parsing here describe the position of the complementizer in different sentences with respect to where it occurs in the whole sentence.

The lexicon, which supports the system, is described below. It contains three thousand words. The lexicon described here is a data structure, which represents a lexeme. Each lexeme has four zones, where each zone corresponds to a slot in the entry zone. The four zones are as follows:

1. Grammatical zone
2. Morphological zone
3. Semantic zone
4. Pragmatic zone

The grammatical zone contains the parts of speech of the word. Consider the following examples:

avan	<Noun><Pronoun>
vannu	<Verb>
karutta	<adj>

In the morphological zone the lexemes has a relevant morphological tag with respect to the grammatical zone. For example, if the grammatical zone is verb then the morphological zone contains the tense and the subcategory transitive/intransitive. If the grammatical zone is noun then the morphological zone contains the case inflexions. The example is as follows.

po:yi	<Verb><trans><pst>
avan	<Noun><Pronoun><NOM>

The semantic zone gives the information regarding meaning, animate-inanimate etc.
avan <Noun><Pronoun><NOM><+animate><+human>

In the last zone the pragmatic zone the other informations like number, gender, person are given.
avan <Noun><Pronoun><NOM><m><s><third><+human><+animate>
The lexicon is alphabetically ordered with respect to the root and tagged with parts of speech.

The parser implemented here does only partial parsing and the anaphora resolution module works on this parser. The lexicon outlined in this chapter is not exhaustive and since anaphora resolution does not require an exhaustive lexicon the present type functions well. This type of lexicon takes large memory space, hence can not handle large data bases.

kutti	sku:lil	vanna	ka:ryam	adhyapakan
child	school-loc	come-RP	matter	teacher

polisino:tu	paRannu	pakse	po:lisu	visvasiccilla	ennu	patranal
police-inst	say-pst	but	police	believe-neg	compl	newspaper

paRannu.
say-pst

(The teacher told the police the fact of the child's coming to the school which the news papers say that the police did not believe.)

The parser output of is as follows.

kutti	<N><NOM><subj><n><s><third><+human><+animate><+count>
sku:lil	<N><LOC><obj><n><s><third><-human><-animate><+count>
vanna	<RP>
karyam	<N><NOM><obj><n><s><third><-human><-animate><-count>
adhyapakan	<N><NOM><subj><m><s><third><+human><+animate><+count>
poli:sinotu	<N><INST><obj><n><s><third><+human><+animate><+count>
paRannu	<V><pst><tran>
po:lisu	<N><NOM><subj><n><s><third><+human><+animate><+count>
visvasiccilla.	<V><neg><pst><intran>
ennu	<compl>
patRannal	<N><NOM><subj><n><pl><third><-human><-animate><+count>
paRannu	<V><pst><tran>

The clauses are:

kutti sku:lil vanna ka:ryam
adhya:pakan poli:sino:tu paRannu
poli:su visvasiccilla
patRannal paRannu

Anaphora resolution system

The algorithm for identifying the antecedent for each anaphor is given below. The input to the anaphora resolution component is the parsed output from the parser.

```
I   1. Create a list of words of S with NPs, clause and immediate clause identified
       (From the parser)
    2. For each pronoun P in S:
       (a) if (P is poss) then
           (i) Select the NP which is the subject of the clause in which P occurs.
           (ii) If P precedes NP in S
```

```
                then STOP and RETURN.
          (b)  else/*(case(p)!=poss)*/
               (i)      identify the immediate clause(s) of the clause in which the pronoun
                 occurs.
           (ii)  if one immediate clause is identified then
                  (a)    select the NP which is the subject of this clause.
                  (b)    if P precedes NP in S then STOP and RETURN
            (iii) if more than one immediate clause then
                   (a)  select the NP which is the subject of the clause preceding the
                        clause containing    P
                 else/*no immediate clause*/
                    antecedent lies outside the S
3.If P and NP do not agree in number, gender, person, then STOP and RETURN
4.The NP is the antecedent of P.
II.  1.  Create a list of words of S with NPs, clause and immediate clause identified
         (From the parser)
      2.  For each reflexive R in S
         (a)  if R is a non-emphatic reflexive then
               (i)    identify the immediate clause of the clause in which R occurs.
              (ii)   select  the NP which is the subject of the immediate clause.
         (b)  if $R_1$ is emphatic or possessive then
               (i)select the NP which is the subject of the clause in which $R_1$ occurs.
     3. The NP is the antecedent of R.
       4. For each reciprocal anaphor R' in S
         (i)select the NP which is the subject of the clause in which R' occurs.
          (ii)if R' precedes the NP.
          The STOP and RETURN
               (i)if NP is not plural
                then STOP and RETURN.
      5. The NP is the antecedent of R'.
      6.For each distributive D in S
        (i)select the NP, which is the subject of the clause in which R occurs.
         (ii)if D precede NP
         then STOP and RETURN
       (iii)if NP is not plural
          thenSTOP and RETURN.
     7. The NP is the antecedent of D.
III. 1. Create a list of words of S with NPs, clause and immediate clause identified
         (From the parser)
     2. Identify the one-pronoun in the list.
     3. Identify the NPs in the IC
       (i)Identify the non-subject NP identified
          (a)if one-pronoun is [+c]
                (i)if not [+count]
                 mark the NP as NON-ANT
         (b)if one-pronoun is[-C]
```

```
                  (i)if not [-animate]
                   (ii)if not [-count]
                    mark the NP as NON-ANT.
            © if one-pronoun is [+/-C]
                    (i)if not SUBJ
                      mark the NP as ANT
      4.The NP(or NP)marked as ANT in (3) is the antecedent of one-pronoun.
III. 1. Create a list of words of S with NPs, clause and immediate clause identified
          (From the parser)
      2. if const(Q')<const(Q") then /*backward*/
      (i). Identify the types of constituents which is in Q" but not in Q'
        (ii)if identified const contains object STOP and RETURN
          (iii)modify Q' by adding the identified constituents in the appropriate slot of Q'
     3. if const(Q')>const(Q") then /*forward*/
      (i). Identify the types of constituents which is in Q' but not in Q"
       (ii)modify Q" by adding the identified constituents in the appropriate slot of Q"
    4. if const(Q')=const(Q") then /*Not a Gap*/
          (i). STOP and RETURN.
   5. Q' and Q" are parallel structures which constitute Q.
IV.1. 1. Create a list of words of S with NPs, clause and immediate clause identified
          (From the parser)
      2. Identify the question words in the Q.
      3. Identify the focus word.
      4. For each W in Q
          if the identical type does not occur in R
                if W is the subj in Q.
                    change W to W' and add to R( by subject change rules)
          else
                    if W is the q-word in Q
                        change W to W'.
                 else add to R.
```

When we use the above algorithm for Hindi data, it yielded a success rate of 82%. The pronoun and reflexive rules for Hindi discovered earlier also show that the modification required to accommodate Hindi is minor. With the modification the system gave the same result as that of Malayalam. The modification required to accommodate Hindi anaphors are in the case of non-emphatic reflexives and in the case of possessive pronoun. As far as the emphatic reflexive is concerned, there is no need for modifying the algorithm. The rule for emphatic reflexive holds good for non-emphatic reflexive. In the case of the possessive pronoun the additional information required to accommodate Hindi is to accept the NP which immediately precedes the possessive pronoun as the antecedent in the case of single clause construction.

Conclusion

This work has developed VASISTH a system for resolving anaphora in two Indian languages, namely, Malayalam and Hindi. Computational grammars of both the languages for the specific purpose are developed.

Coming to anaphora resolution the algorithms for reflexives, we find that the algorithms for one-pronoun, reciprocal and distributive do not require any change. As far as the elliptical constructions are concerned, Malayalam has two types: wh-constructions and que-constructions, whereas Hindi has only one type: wh-construction. The grammars of such constructions of both the languages are quite similar, which makes the same algorithm work for both the languages. Gapped constructions too show great similarity; however we would like to draw attention to just a single difference between the grammars, which is with respect to backward gapping in Hindi. Some changes were necessary for pronoun resolution as the rules are different in both cases. As for the remaining ones, the same algorithm works.

The lexicon for Hindi required minor changes, which is not unexpected. Hindi has the same marker for different case inflections like "se", and "ko" etc. Hindi, unlike Malayalam, is a language that shows agreement; so the lexicon had to be modified to capture it.

Although at the moment VASISTH can handle two languages, we strongly believe that this system can be used without major modification for all Indo-Aryan, Indo-Dravidian and Indic family of languages, in general for all morphologically rich languages.

The limitation of this work is that it uses only syntactic knowledge and does not use any world knowledge for resolution. Even without this, the success is not discouraging; we have 93% success in pronoun and one-pronoun resolution and interpretation of gaps, and 96% success in the case of reflexive and ellipsis both. The resolution which basically handles sentences, and very short discourses, can hopefully be easily extended to cover longer discourses and corpora, but this requires to be examined.

The system works with high degree of success in the case of Malayalam. Evaluation shows a success rate of 82% in the case of Hindi when there was no modification. This can be extended to other Indian languages in particular and to morphologically rich languages in general.

Bibliography

Allen, J. (1987). Natural Language Understanding. *Benjamin/cumming Publishing company*, Inc. California.

Baldwin, Breck. (1997). "CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources", *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 38-45, Spain.

Carbonell, J and R Brown. (1988). "Anaphora Resolution: A Multistrategy Approach", *Proceedings of the 12th International Conference on Computational Linguistics*, 96-101.

Hobbs, Jerry. (1978). "Resolving pronoun references", *Lingua*, 44. 311-338.

Kennedy, Christopher and Branimir Boguraev. (1996). “Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser”, *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, 113-118, Denmark.
Lappin, Shalom and M Mccord. (1990). “Anaphora Resolution in Slot grammar”, *Computational Linguistics,* 16, 4, 197-210.

Lappin, Shalom and Herbert Leass. (1994). “An Algorithm for Pronominal Anaphora Resolution”, *Computational Linguistics*, 20, 4, 535-561.

Mitkov, Ruslan. (1997). “Factors in Anaphora Resolution: They are not the only Things That Matter. A Case Study Based on Two Different Approaches”, *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 14-21, Spain.

Mitkov, Ruslan. (1997). “How Far are We from (semi-) Automatic Annotation of Anaphoric Links in Corpora?”, *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 82-87, Spain.

Mitkov, Ruslan. (1998). “Robust Pronoun Resolution with Limited Knowledge”, *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference, (forthcoming*), Canada.

Mitkov, Ruslan. ”.(Forthcoming),“Evaluating Anaphora Resolution Approaches”.

Mitkov, Ruslan and Malgorzata Stys. (1997). “Robust Reference Resolution with Limited Knowledge: High Precision Genre-Specific Approach for English and Polish”, *Proceedings of the International Conference “Recent Advances in Natural Language Proceeding”(RANLP'97)*, 74-81, Bulgaria.

Mitkov, Ruslan, Lamia Belguith and Malgorzata Stys. (1998). “Multilingual Robust Anaphora Resolution”, *Proceedings of the Third International Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*, 7-16, Spain.

Sobha L, B.N. Patnaik. (1998). “An Algorithm for Pronoun and Reflexive Resolution in Malayalam”, *Proceedings of the International Conference on Computational Linguistics, Speech and Document processing*, C63-66.

Sobha L, B.N.Patnaik. (1999). “One-pronoun resolution in Malayalam”, Indian Linguistics

Sobha L, (1999) “Anaphora Resolution In Malayalam and Hindi” Unpublished Doctoral dissertation. Mahatma Gandhi University, Kottayam , Kerala.