Anaphors in Sanskrit

Girish Nath Jha
Center for Sanskrit Studies
J.N.U. New Delhi-67, India
girishjha@gmail.com

Sobha L
AU-KBC Research Centre
Anna University, Chennai-44, India
sobhanair@yahoo.com

Diwakar Mishra
Center for Sanskrit Studies
J.N.U. New Delhi-67,India
diwakarmishra@gmail.com

Surjit Kumar Singh
Center for Sanskrit Studies
J.N.U. New Delhi-67, India
surjit.jnu@gmail.com

Pravin Pralayankar
AU-KBC Research Centre
Chennai-44, India
praveen@au-kbc.org

Abstract

Research in building robulst NLP systems with ambiguity resolution techniques has gained momentum in recent years. In particular, the anaphora resolution initiatives have reached unpreceented heights in last 10 years or so. Mitkov et. al (2001) have reported both rule based knowledge based approaches and machine learning based 'knowledge poor' approaches in an ACL issue devoted to this subject. Mitkov (2001 a) has also presented outstanding issues and challenges in this area. Johansson ed. (2007) reports the latest developments in this area of research and development.

Indian languages in general, and Sanskrit in particular have not been profusely worked upon from these perspectives. Barring a notable exception (Sobha 2007), Sanskrit anaphors have been rarely looked upon from a computational perspective. The case of Sanskrit has been more severe due to two reasons - a virtual absence of annotated corpora made it impossible for corpus based machine learning approaches and a poor understanding of Pāṇini's grammar from computational perspective has made it difficult to apply rule based approaches. While some work on Indic languages like those of Hock (1991), Davison (2006) have looked at diverse syntactic issues often not excluding anaphora, Shapiro (2003) has focused on lexical anaphors and pronouns in the languages of the subcontinent. Sobha et al (1998, 1999-a&b, 2007), Murthy et al (2005) have looked into the anaphora cases for some Indian languages in great detail and in particular for Sanskrit in their most recent paper (2007) as mentioned above.

The authors in this paper are looking at the problem in a broader perspective. Since no effort has been made at comprehensive documentation and classification of Sanskrit anaphora, this is the primary focus of the present study. Similar to Soon, Ng and Lim (2001), the anaphora resolution presented here is proposed to be a part of the larger NLP system called Sanskrit Analysis System parts of which have been developed by the principal author and his research students at the Sanskrit Center of Jawaharlal Nehru University, New Delhi (Jha et al 04,05,06,07,08). The paper has the following major sections –

- **Sanskrit and its linguistic tradition**: This section gives a brief background of Sanskrit language and its linguistic tradition for the benefit of general readers

- **Anaphors in Sanskrit**: This section is an attempt at discussing and classifying anaphora and anaphora-like cases occurring in a wide variety of Sanskrit prose texts – those from the authors like Subandhu

(*Vāsavadatta),* Bāṇabhaṭṭa(*Kādambarī, Harṣacarita*), Daṇḍī (*Daśakumāracarita*), Ambikā Dutt Vyāsa (*Śivarājavijaya*) and popular didactic prose texts like *Pañcatantra* and *Hitopadeśa*. Though the focus of this paper is the classical Sanskrit prose, examples from some popular poetic texts like *Bhagvadgītā* and *Rāmāyaṇa* have also been studied. The authors have looked at the above textual data to arrive at a classification of lexical, sentential and discourse anaphora in Sanskrit.

- **Anaphora handling in Sanskrit intellectual tradition**: this section presents three major śāstrāic (scientific) traditions in India which have insights on this subject – the vyākaraṇa (grammar), the navya-nyāya (logic) and mīmāṁsā (interpretation) schools. An attempt has been made to understand what solutions have been provided to handle such ambiguity at various levels of natural language.

- **Computational Framework**: In this section, available computational models for anaphora resolution for Indic languages have been examined and a new model has been attempted after incorporating inputs from the traditions of *vyākaraṇa* (grammar), *mīmāṁsā* (interpretation), and *nyāya* (logic). The anaphora resolution presented here is a subset of a larger web based Sanskrit Analysis System implemented in Java (http://sanskrit.jnu.ac.in) and comprising of the following components –

    - Sandhi analysis (http://sanskrit.jnu.ac.in/sandhi/viccheda.jsp)

    - Morph analysis for nouns (http://sanskrit.jnu.ac.in/subanta/rsubanta.jsp)

    - Morph analysis for verbs (http://sanskrit.jnu.ac.in/tanalyzer/tanalyze.jsp)

    - POS tagger (http://sanskrit.jnu.ac.in/post/post.jsp)

    - Gender recognition system (http://sanskrit.jnu.ac.in/grass/analyze.jsp)

    - Semantic class identification based on amarakosha (http://sanskrit.jnu.ac.in/amara/index.html)

    - Karaka analysis (http://sanskrit.jnu.ac.in/karaka/analyzer.jsp

and depends on the information provided by the above components like – morphosyntactic labels, morphological information, case and gender information etc.

1. Sanskrit and its linguistic tradition

Sanskrit is the oldest documented language of the Indo-European family. *ṛgveda* (2000 BCE) is the oldest text of this family contains a sophisticated use of the pre-Pāṇinian variety also called *vaidikī*. Pāṇini variously calls his mother tongue as *bhāṣā* or *laukikī* . His grammar has two sets of rules – for *vaidikī* (variety used in the vedas) and for *laukikī* (variety used by the common people). The term 'Sanskrit' (meaning 'refined') is given to the standard form of laukikī which emerged after Pāṇini's grammar (700 BCE).

Many indologists have studied the evolution of Sanskrit in three phases –

1.1 Old Sanskrit (Vedic)

Vedic (*vaidikī*) is the oldest extant form of Sanskrit or Old Indo Aryan (OIA). It roughly dates back to 2000 BCE and includes the four Vedas (***ṛg-veda, sāma-veda, yajur-veda, atharva-veda***), their ***pāṭhas*** (structured readings) and ***samhitā*** traditions – for example *taittirīya* and *maitrāyiṇi samhitās* of *kṛṣṇa yajurveda*, and *vājasaneyī samhitā* of *śukla-yajurveda*. Based on linguistic similarities, some scholars have also included the later pre-classical phase of Sanskrit including *brāhmaṇas, āraṇyakas, upaniṣadas* and *kalpasūtras* as vedic

1.2 Middle Sanskrit (pre-classical)

The middle Sanskrit consists of ***brāhmaṇas, āraṇyakas, upaniṣadas*** belonging to the saṁhitas, ***kalpasūtras*** (kalpa vedāṅga), and the six ***vedāṅgas*** or scientific disciplines required to be studied for understanding Vedas. Four out of these dealt with linguistics -

- *śikṣā* (pronunciation)
- *vyākaraṇa* (grammar)
- *nirukta* (etymology)
- *chanda* (meter)
- *jyotiśa* (astronomy)
- *kalpa* (ceremonial)

Besides these, many indices and lists were prepared for explaning the vedic verses. These were called *pariśiṣṭa* (appendices explaining *sūtras*) and *anukramaṇī* (lists containing order of verses and information on organization of vedic texts)

1.3 Classical Sanskrit

The classical Sanskrit includes the epics (***mahabhārata*** and ***rāmāyaṇa***), 18 ***purāṇas***, and and a huge body of ***sāhitya*** (literature) many of which laid the foundation of indological studies in the west.

1.4 Purpose of Linguistic studies in India

The ancient Indian scholars were pre-occupied with linguistic studies for two basic reasons - to maintain texts of oral tradition, and to defend the Vedic knowledge. As mentioned above, of the six ancillary disciplines required to understand Vedas, four were for linguistic study. Kapoor (1993) has divided the Indian linguistic tradition in four phases -

Phase I: earliest times up to Panini (2000 – 700 BCE)

Speculations in śruti texts, four of the six vedangas (vyākaraṇa, chanda, nirukta, śikṣā), work of Yāska, ṛk prātiśākhya, ācāryas mentioned by Pāṇini.

Phase II: Pāṇini to Ānandavardhana (700 BCE - 9th CE)

aṣṭādhyāyī of Pāṇini, *vārttika* of Kātyayana, *mahābhāṣya* of Patañjali, *mīmāṁsāsūtra* of Jaimini, *vākyapadīya* of Bhartṛhari, works on poetics from Bharata up to Ānandavardhana.

Phase III: Rāmacandra to Nāgeśa Bhaṭṭa (11th CE to 18th CE)

This phase ncludes pedagogical grammars based on Pāṇini's grammar, investigations into principles of grammar and also attempts to apply Pāṇinian model to describe other languages.

Phase IV: Franz Kielhorn onwards

This phase includes modern textual interpretations of language, works of Kielhorn, Bhandarkar, Carudev Shastri, Katre, Dandekar, among many others.

2 Anaphors in Sanskrit

Sanskrit anaphors can be classified in two broad categories – anaphor proper and anaphor-like cases. Sobha et al (2007) have taken the first category and considered only the pronominals like *tat* (pronominal), *yat* (corr pronoun), *sva, ātman* (reflexive), *parasparam, anyonyam* (reciprocal) and their inflected forms as anaphors.

Pāṇini in his sūtra *sarvādīni sarvanāmāni* lists the following pronouns (*sarvanāma*) –

sarva, viśva, ubha, ubhaya, ḍatara, ḍatama, anya, anyatara, itara, tvat, tva, nema, sama, sima; pūrva, apara, avara, dakṣiṇa, uttara, apara, adhara (when not used as noun); *sva* (if not used for family or wealth); *antara; tyad, tad, yad, etad, idam, adas, eka, dvi, yuṣmad, asmad, bhavatu, kim* (35)

These can be categorized as follows –

- sarvādi (14) including two sets of ḍatara and ḍatama which are suffixes to form comparatives and superlatives
- *pūrva, apara, avara, dakṣiṇa, uttara, apara, adhara* (when not used as noun)
- *sva* (if not used for family or wealth), example - *hariḥ svān vedārthaṁ avedayat*in this case it is used as a noun in the sense of family or wealth, *ātman* (though not a pronoun, can come as reflexive) anaphor
- *antara* (noun if used in the sense of under garments)
- *tyad, tad, yad, etad, idam, adas, eka, dvi, yuṣmad, asmad, bhavatu, kim*

Chandrashekar (2007) classifies pronouns in Sanskrit as follows –

SN	Sarva Nāman (Pronoun Other, with gender, number, and declensional sub-tags) (e.g., *anyaḥ, aparā)*
SNU	Sarva Nāman Uttama (Pronoun First Person, number, and declensional sub-tags) (e.g., *asmad)*
SNM	Sarva Nāman Madhyama (Pronoun Second Person, number, and declensional sub-tags) (e.g., *tvad*)
SNA	Sarva Nāman Ātman (Pronoun Reflexive, with or without gender, number, and declensional sub-tags) (e.g., *nijaḥ, svasya)*
SNN	Sarva Nāman Nirdeśātmaka (Pronoun Demonstrative, with gender, number, and declensional sub-tags) (e.g., *idam, saḥ)*
SNP	Sarva Nāman Prāśnārthika (Pronoun Interrogative, with gender, number, and declensional sub-tags) (e.g., *kim, kad)*
SNS	Sarva Nāman Sāmbandhika (Pronoun Relative, with gender, number, and declensional sub-tags) (e.g., *yaḥ, yā)*

2.1 Classification of anaphors in Sanskrit

Here we try to deal with anaphora and similar elements in Sanskrit language usage. To see it in broader perspective, the cases of this are firstly divided into ‘anaphora proper’ and ‘anaphora like’ cases.

‘Anaphora proper’ can be divided into four categories – pronominal, reflexive, reciprocal and correlative.

a) pronominal anaphors

where there is one or more pronouns and they refer to some other noun or pronoun.

1.
[*rāmaḥ[i] gṛham āgataḥ*]**MC** [*saḥ[i]*
Rama[i] to home came he[i]
ca **CONJ** *mayā dṛṣṭaḥ*]**MC**
and by me was seen

2.
[*rāmaḥ[i] gṛham āgataḥ*]**MC** [*aham*
Rama[i] to home came I
[*ca*] **CONJ** *tam[i] apaśyam*]**MC**
and him[i] saw

In the first part of the next sentence (3a), an NP antecedent precedes the pronominal anaphor in the same clause. The latter part (3b) is an example of correlative anaphor and will be discussed later

3. a.

sattva-anurūpā[i] *sarvasya[i]*

self-nature[adj.f.nom.sg]i all[pro.m.gen.sg]i

śraddhā *bhavati*

faith[f.nom.sg] be[pres.III.sg]

bhārata

arjuna[m.nom.sg]

3.b.

śraddhā-mayaḥ *ayam*

faith-ful[adj.m.nom.sg] this[pro.m.nom.sg]

puruṣaḥ *yaḥ*

person[m.nom.sg]j who[pro.m.nom.sg]j

yat-

which[pro.m.nom.sg]k

śraddhaḥ

faithful[adj.m.nom.sg]

saḥ *eva*

he[pro.m.nom.sg]j same[indecl]

saḥ

he[pro.m.nom.sg]k

Both 3.a and 3.b. are from (Gītā-17.3) and read as - *Arjuna! each person’s[i] faith is according to his/her own[i] nature. A person[j] having faith is identical with the person/thing he/she[j] has faith in*

4)
ātma-aupamyena

self[pro]-comparison[n.ins.sg]i

sarvatra *samam*

everywhere[indecl] equally[indecl]

paśyati *yaḥ*

see[pres.III.sg]who[pro.m.nom.sg]i

arjuna *sukham*

arjuna[m.voc.sg] happiness[n.acc.sg]

vā *yadi* *vā*

or[indecl] if[indecl] or[indecl]

duḥkham *saḥ[i]*

grief[n.acc.sg] he[pro.m.nom.sg]

yogī *paramaḥ*

yogī[m.nom.sg] absolute[adj.m.nom.sg]

mataḥ

regarded[adj.m.nom.sg]

Arjuna! One who[i] sees happiness and sorrow equally as compared with oneself[i], he [i] is regarded as absolute yogī.(Gītā-6.32)

b) Reflexive anaphors

where there are reflexive pronouns or words such as *sva, svayam, svayameva, ātma, ātmanā* and other forms of these and they refer to some noun or pronoun in its left or right context.

For example

(5)
rāmaḥ[i] *sva[i]-gṛham* *gacchati*
Rama[i] self[i]-home goes

(6)
aham[i] *ātmānaṁ[i]* *prati*I[i]
I[i] self[i] about
sacetaḥ *asmi*
conscious am

c) **Reciprocal anaphors**

This is marked by the use of words like *parasparam* as in the following example –

7)
kṛṣṇaḥ pārthaḥ ca
kṛṣṇa[i] pārtha[i] and
parasparam paśyataḥ
one another[i] see [pres.III.dual]

d) **Correlative anaphors**

where there is some correlative pronoun (generally demonstrative) which binds an NP headed by a relative pronoun as its antecedent. As an example, we will reproduce sentence 3.b. as 8 –

8)

śraddhā-mayaḥ *ayam*

faith-ful[adj.m.nom.sg] this[pro.m.nom.sg]

puruṣaḥ *yaḥ*

person[m.nom.sg]j who[pro.m.nom.sg]j

yat-

which[pro.m.nom.sg]k

śraddhaḥ

faithful[adj.m.nom.sg]

saḥ *eva*

he[pro.m.nom.sg]j same[indecl]

saḥ

he[pro.m.nom.sg]k

(9)
yat yat[i] *karomi* *śiva[j]*
whatever[i] I do o Shiva
tat tat[i] *sarvam[i]* *tava[j]*
that[i] all[i] your
eva *ārādhanam*
only worship

10)
yasya *asti*

who[pro.m.gen.sg]i be[pres.III.sg]

vittam *saḥ*

wealth[n.nom.sg] he[pro.m.nom.sg]i

naraḥ *kulīnaḥ,*

man[m.nom.sg]i well-born[adj.m.nom.sg]

saḥ *paṇḍitaḥ*

he[pro.m.nom.sg]i learned[adj.m.nom.sg]

saḥ *guṇavān*

he[pro.m.nom.sg]i virtuous[adj.m.nom.sg]

guṇajñaḥ

expert-critic[adj.m.nom.sg]

saḥ *eva*

he[pro.m.nom.sg]i only[indecl]

vaktā *saḥ*

speaker[adj.m.nom.sg] he[pro.m.nom.sg]i

ca *darśanīyaḥ*

and[indecl] good-looking[adj.m.nom.sg]

sarve *guṇāḥ*

all[pro.m.nom.pl] virtue[m.nom.pl]

kāñcanam *āśrayanti*

gold[m.acc.sg] reside[pres.III.pl]

one who[i] has money, he[i] is well-born, (he[i]) is learned, (he[i]) is virtuous and expert critic, (he[i]) is good speaker, (he[i]) is good looking as all the virtues lie in the gold. (subhāṣitam)

11)
ye[i] api ca mantra-karkaśāḥ tantra-

who[pro.adj.m.nom.pl]i

kartāraḥ *śukra-āṅgirasa-viśālākṣa-*

bāhudantiputra-parāśara-prabhṛtayaḥ

taiḥ[i] kim ari-ṣaḍ-vargaḥ jitaḥ, kṛtam vā

he[pro.m.instr.pl]i

taiḥ[i] śāstra-anuṣṭhānam?

he[pro.m.ins.pl]i

taiḥ[i] api hi prārabdheṣu kāryeṣu dṛṣṭe

he[pro.m.ins.pl]i *siddhi-asiddhī*

(Daśakumāracarita-viśrutacarita- p.261)

wosoever[i] - the polity experts[i/j] and system profounders[i/j] śukra[i/j] āṅgirasa[i/j] viśālākṣa[i/j] bāhudantiputra[i/j] parāśara[i/j] etc. - have been there, have they[i] overcome the six enemies, or did they[i] observe the p recepts? They[i] also have seen the success and failure in their[i] tasks started.

3 Anaphora handling in Sanskrit intellectual tradition

From the instances above, it becomes amply clear that Sanskrit literature abounds in anaphora cases. It is then very surprising that the Indian śāstraic tradition while emphasizing on correct textual interpretation, has ignored anaphor resolution largely. One possible reason for this could be the fact that most of the ancient Sanskrit literature was in the oral tradition with short and racy verses with rigid grammatical agreements which made anaphora binding obvious. It is only in the later period when long prose antholologies started emerging, that this problem would have become apparent across sentence boundaries. However, by this time the focus in linguistic tradition may have shifted to the philosophy of grammar and not grammar per say.

3.1 *Vyākaraṇa* (Grammar)

Patanjali's "capability principle" called ***samarthaḥ pada-vidhiḥ*** expounded in the context of compounds can be extended to anaphora. According to this principle –

I: Two words W1 and W2 can compound iff they have *sāmarthya*

II Any two entity has *sāmarthya* of forming a relationship iff they have

- mutual *ākāṅkṣā* (expectancy) and
- *yogyatā* (compatibility) and
- *sannidhi* (contiguousness) and
- a common reference

This condition of compatibility is also applicable in the case of anaphor binding. Where the grammatical categories are not sufficient to find the exact match or bind for anaphors, the expectancy and compatibility are the only ways to resolve anaphora. Even if the similarity in grammatical features allows a certain anaphoric relation, the condition of compatibility is a must.

Samāsa (compound) is applied only in two cases, *ekārthībhāva* (joining two meanings to one meaning) and *samāhāra* (collective meaning). If it seems that more words are compounded, there it is also according to these rules. For example, if three words are compounded, it means, first two are compounded, then third is compounded with the compound of the other two- [(w1⇔w2)⇔w3] or [w1⇔(w2⇔w3)]. This condition is also applicable in the case of anaphora. Example of *ekārthībhāva* is when a pronoun of dual number is anaphor and its antecedents are two different wordsof singular number. Here problem comes that one word cannot be co-referential with two words and the number of anaphor does not match with either of the antecedents. This is solved by *ekārthībhāva* that the two singular words together constitute one meaning with dual number. Thus it becomes compatible of being co-referrential with one word of dual number. For example, *rāmaḥ[i] śyāmaḥ[j] ca āgatau, tau[i/j] adhunā nivartataḥ*. Example of *samāhāra* or collective meaning is where one antecedent or anaphor binds more than one anaphors or antecedents. All together constitute a collective meaning and that meaning is compatible with other of anaphora and antecedent. Such as in *yasyāsti vittaṁ saḥ naraḥ kulīnaḥ* ....

As the exceptional rule (*niravakāśo vidiḥ apavādaḥ*) is dominant rule because it has no other scope. This is also similar in the case of anaphora that in the case of many anaphors and many antecedents, if one anaphor can bind more than one antecedents and another can bind only one of them, then due to having no other scope, the second anaphor is more likely to bind that particular antecedent. For example, *rāmaḥ*_m *phalam*_n *ānayat, tena*_m/n *tat*_n *khādyate*. Here pronoun *tena* can bind both the nouns because it is the same form in masculine and neuter, but *tat* is neuter pronoun and it can bind only *phalam* and not any other antecedent. So due to no other scope, *tat* will bind *phalam* and *tena* which has two possibilities of antecedent, will bind the other antecedent *rāmaḥ*.

3.2 *Nyāya* (logic) theory

If a word refers to another, the relation is directed and exists between those two words only, not with their synonyms.

- a→b ≠ a→b’
- a→b ≠ b→a

co-referentiality is of many types-

a) Noun with noun (3 types)

- two nouns with same inflectional suffix as *rāmo jāmadagnyaḥ*
- two nouns in a compound as *kṛṣṇasarpaḥ*
- nouns which are co-referential by other causes

For example - *nīlam utpalam.* Here *nīla* and *utpala* are co-referential, but if some synonym of the same like *paṅkaja* or *kamala* comes in the same sentence, and it refers to the first one, then *utpala* and *paṅkaja* are co-referential

if *nīlam →utpalam*
and nīlam →pankajam
then *utpalam →pankajam*

b) Verb root with nominal base

In example *sukham āste* (he/she lives happily) *sukham* is a nominal base and acts as adverb. This adverb becomes co-referential with verb root *ās.*

c) Inflectional suffix with nominal base

Here inflectional suffix should be understood as suffix in general, because the examples and explanations of this case include verb inflectional suffixes and nominal derivation suffixes also. Inflectional verb suffix in active voice denotes or refers to the agent. On the other hand, the agent is denoted by a nominal base in the same sentence. Thus both are co-referential. Also the meaning of *kṛt* (noun derivational) suffixes is the agent (*kartari kṛt*). Thus these suffixes also become co-referential with the agent which is a nominal base.

3.3 *Mīmāṁsā* theories

A rule of mīmāṁsā is quoted here in the commentary that the combination of the words connected with *yat* or words meaning 'which' is called firstness, definiteness and recallable. The combination of words connected with *tat* ('that') is called laterness, difinableness and predicatable. Without stating recallable, predicate should not be used. This rule tells that the antecedents of relative-correlative anaphora are always found in the left of *that*.

The theories of determining whole-part relationship (six scales) may be useful for adapting the rules of anaphora resolution. Six scales or *pramāṇa*s of *viniyoga-vidhi* are- *śruti, liṅga, vākya, prakaraṇa, sthāna* and *samākhyā*. Among these, the prior is stronger or dominating over the latter. Whenever prior is not applicable, then only the latter is applied for decision.

Śruti- where one thing is literally stated as the part of other action or anything. *nirapekṣo ravaḥ śrutiḥ* or the absolute utterance is *śruti* scale of *viniyoga vidhi*.

Liṅga- where one thing is inferred as the part of the other through the implication of the other word closely related. *śabda-sāmarthyaṁ liṅgam* or power of word of denoting the meaning that is denotation itself- *sāmarthyaṁ rūḍhireva*

Vākya- where one thing is known as the part of the other on the basis of their being together uttered in a sentence.

Prakaraṇa- in a context in different sentences, when two things have expectancy of each other, i.e., one sentence has verbalization of part or *upakāraka* or *aṅga* only and not of whole, another has verbalization or utterance of whole or *upakārya* or *aṅgī* only and not of part, by the expectancy they become whole-part of one another.

Sthāna- it means closeness. When one thing is decided as the part of other on the basis of closeness, this scale is called *vākya-pramāṇa*. This is on two bases, closeness in utterance in the text, and other is closeness while performing the action. Text closeness is again of two types. First case is of similar sequence, where wholes are stated with a sequence and parts are stated with same sequence. Then first part is part of first whole, second is of second, and so on. Second case is where part is uttered near the whole.

Samākhyā- when a thing is recognized as the part of the other on the basis of its etymology, this scale is *samākhyā-pramāṇa*. (needs explanation of adaptation).

4 Computational Framework

4.1 Recognition of POS categories

First the grammatical catagories are marked by POS tagger. It will give the information of case, person, number and gender (CPNG). In case of POS failure, the string needs to be sent for sandhi and samāsa (compound) processing, before trying POS again. For example, *sva-gṛhaṁ* in the sentence *rāmaḥ sva-gṛhaṁ gacchati*, may be wrongly marked as Noun or remain untagged as the reflexive *sva* is attached with the noun. Unless this is separated by way of de-compunding or sandhi-splitting, the POS tagger will not recognize it correctly.

Examples –

INPUT: *rāmaḥ sva-gṛhaṁ gacchati*

OUTPUT1: *rāmaḥ*[N_MAS_NOM_SG] *sva-gṛhaṁ*[??_NEUT_ACC_SG] *gacchati* [V_PRES_III_SG]

The label '??' in a token requires it to be sent for sandhi and samasa processing

OUTPUT12: *svam gṛhaṁ*

OUTPUT13: *svam*[PRON_REFLX_NEUT_ACC_SG] *gṛhaṁ*[N_NEUT_ACC_SG]

4.2 Recognition of sentences

Sentence has been variously defined in Sanskrit. Unlike other written languages where it can be identified by punctuation, in Sanskrit there may be no punctuations in the text. Therefore we may have to go for identifying main verbs in the string to identify a sentence. The typical definition is *eka tiṅ vākyam* (one verb sentence)

The information from POS tagger will be used by *kāraka* (case) analyzer to identify the verb, *kārakas* and thus sentences, because there is no rigid convention of punctuations after each sentence. The sentence can be recognized by *daṇḍa* or by the *kāraka* web (the web of verb and its objects or *kārakas*).

Strategy

- if the category is detected as reflexive, then the antecedent will be in the same sentence
- if a relative is found, then search for correlative in the next clause/s or sentence/s
- in any other case of pronoun, look for antecedent in the prevous sentence/s

4.3 Marking possible antecedents and anaphors

- The nouns in a sentence will be marked as possible antecedents because a noun will never be an anaphor in the category of anaphors we are exploring (pronominal anaphors). Also it is not necessary for a noun to be an antecedent (when there is no anaphor or zero anaphor), so it is said to be a possible antecedent. In such cases, if pronouns are found elsewhere in the sentence, then such nouns will be considered as potential binding candidates

- Relative pronouns are always antecedents. Therefore they are marked as antecedents. In these cases, the anaphora hunting goes to the next clause(s)/sentence(s). This can lead to a whole discourse situation or at least two analyzable sentences (one verb with the *kāraka* web). If relative pronoun exists, other third person pronouns will be marked as anaphor because the relative pronouns (antecedents) must have correlative pronoun(s) to complete the sense.

- The other pronouns of third person are more likely to be anaphor, may be sentential or cross-sentential. So these pronouns will be marked as possible anaphors.

- In case of reflexive, the antecedent noun must be in the same sentence, most likely, preceding the anaphor.

- Second person pronouns will come as anaphors very rarely. They are mostly antecedents. So it will be marked as possible antecedent and also possible anaphor (when there is a noun in vocative case). If the noun antecedent is in vocative case, then 2nd person pronouns can be anaphors as in the following sentence –

  yat yat[i] karomi ***śiva*****[j VOC]** *tat tat[i] sarvam[i]* ***tava*****[j II P.GEN]** *eva ārādhanam* or 'whatever[i] I do **o Shiva[j VOC]**, that[i] all[i] is **your[j II P.GEN]** worship.'

- First person pronouns will not come as anaphors, but only as possible antecedents in cases of reflexive anaphors in the sentence.

4.4 sentences excluded from analysis

Following sentences will be excluded from analysis –

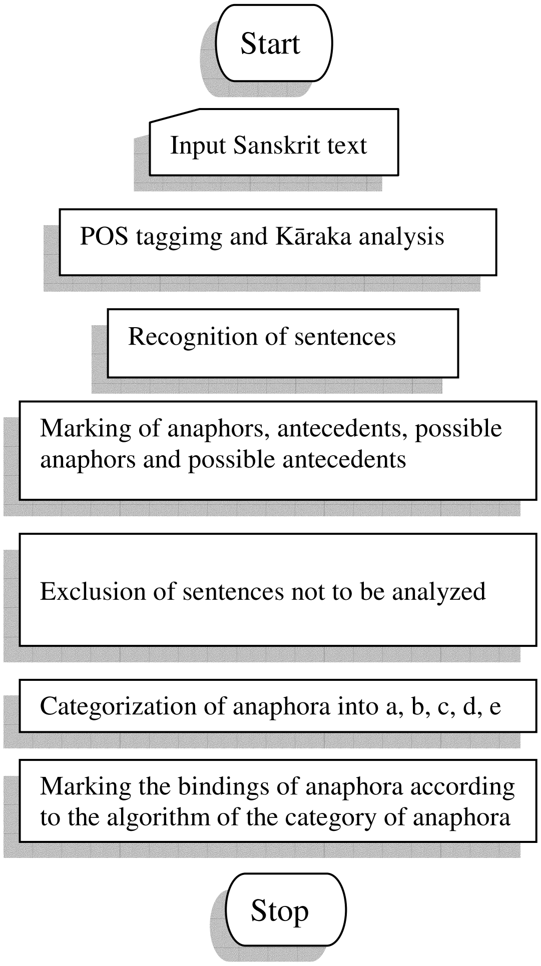
- sentence with no anaphor
- sentence with reflexive anaphor with a valid antecedent (because the binding is obvious in such cases)

4.5 Categorization of problematic anaphoric sentences

The remaining sentences with anaphora cases will be divided in the following five catagories-

a) only one anaphor and zero or one antecedent
b) one antecedent and two or more anaphors
c) two or more antecedents and one anaphor
d) two or more antecedents and equal number of anaphors
e) two or more antecedents and unequal number of anaphors

For analysis of anaphora in these categories, one sentence from left and one sentence from right will be considered because these types of anaphors have their antecedents often in other sentences. In these cases, ‘sentence’ means three sentences together.


Start
Input Sanskrit text
POS taggimg and Kāraka analysis
Recognition of sentences
Marking of anaphors, antecedents, possible anaphors and possible antecedents
Exclusion of sentences not to be analyzed
Categorization of anaphora into a, b, c, d, e
Marking the bindings of anaphora according to the algorithm of the category of anaphora
Stop

4.6 Algorithm for category (a) anaphora binding

with antecedent

- if the anaphor is in II person and there is one antecedent in vocative case, both are bound with each other.

- if the anaphor is in III person and of demonstrative type and the antecedent agrees with it in terms of CPNG, both are bound with each other.

- if the anaphor is in III person and of demonstrative type, and the antecedent is fixed gender and does not agree with anaphor’s CPNG, but if they are compatible, they are bound.

without antecedent

- find the antecedent in the previous sentence which has common CPNG. Vocative case should be considerd as II person. If found then verify by compatibility. Its antecedent can be noun or pronoun

- if there is no antecedent with CPNG agreement then find antecedent with PNG agreement. If found then verify by compatibility.

- if antecedent with PNG agreement not found then check whether some antecedents constitute collective meaning and there is number agreement with anaphor. If found then verify by compatibility

- if antecedent with number agreement not found then find a relative pronoun in the next sentence as antecedent with CPNG agreement. Then verify if it is not bound with some other correlative pronoun. If verified, then bind both

- if not found then find a relative pronoun in the next sentence as antecedent with PNG agreement. It is also to be verified if it is not bound with some other correlative pronoun. If verified, then bind both.

4.7 Algorithm for category (c) anaphora binding

Choose the antecedent and search in the sentence for (possible) anaphora with CPNG agreement.

- if only one such anaphor is found then it is bound with it

- if there are more than one such anaphors, and if they constitute collective or integerated meaning, then all those anaphors bind the antecedent.

- if there are more such anaphors but do not constitute a collective meaning, one or more of them are selected to bind on the basis of compatibility and semantics, which have same kāraka from the view of meaning.

If there is no anaphor with CPNG agreement, then search for anaphors with similar PNG to antecedent

- if only one such anaphor is found then it is bound with it.

- if there are more than one such anaphors, and if they constitute collective or integerated meaning, then those all anaphor bind the antecedent.

    - if there are more such anaphors but do not constitute a collective meaning, one or more of them are selected to bind on the basis of compatibility and semantics, which have same kāraka from the view of meaning.

4.8 Algorithm for category (b) anaphora binding

The anaphor with CPNG agreement binds with the antecedent. If no antecedent with CPNG agreement is found, then the antecedent with PNG agreement gets bound

4.9 Algorithm for category (d) and (e) anaphora binding

First determine which kind of anaphora cases are there in the sentence - regular, reflexive or relative-correlative

- if there is a reflexive pronoun, it will be bound with a competent antecedent (can be noun or pronoun) with matching NG (because reflexive pronoun has no gender distinction) in the same sentence

- if there is only one relative antecedent (pronoun), it will bind to a correlative anaphora with similar CPNG

    - if more than one anaphor with CPNG agreement are found then one of them will be chosen by compatibility and expectancy (semantics)

    - if CPNG match is not found then check for anaphors (correlative pronoun) of only PNG match

    - if more than one anaphor with PNG agreement are found then one of them will be chosen by compatibility and expectancy (semantics).

▪ if all antecedents are relative pronouns, each one will be bound by one independent anaphor

    - rank antecedents and anaphors as 1, 2, 3… and bind them respectively (first with first, second with second and so on) after verifying CPNG then PNG then semantic compatibility

- if antecedents and anaphors are in mixed order, then bind each relative pronoun to the nearest anaphor and verify by similar CPNG or PNG or semantic compatibility

If there are more anaphors remaining, they all are proper anaphora. For them search for CPNG match in remaining possible antecedents (that is more likely to be noun but can be pronoun also). The preferred order of verification would be CPNG > PNG > semantic compatibility

5. Conclusion

The authors have tried to present a wider case study of anaphors in Sanskrit scaning classical Sanskrit from the epics to *śivarājavijaya*. Sanskrit language abounds in anaphors for reasons of precision and brevity in the oral tradition and for elegance in the literary tradition. In this paper we have focused only on pronominal anaphors and have presented a workable solution in terms of an algorithm. The resolution strategy presented here is part of a larger Sanskrit analysis system. Several modules of the SAS have been developed and are in a testing and evaluation phase. The authors would like to have feedback on this paper and would also like to present the non-pronominal anaphors in Sanskrit in ther subsequent endeavours.

References

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

Chandrashekar R. 2007. POS tagging for Sanskrit. Ph.D. dissertation, J.N.U., New Delhi

Davison Alice. 2006. Syntactic effects of correlative clause features in Sanskrit and Hindi/Urdu, University of Iowa

Hobbs Jerry. 1978. Resolving pronoun references. In *Lingua*, 44. 311-338.

Hock Hans Henrich (ed) 1991. Studies in Sanskrit Syntax, Motilal Banarasidass, Delhi

Jha Girish Nath. 2004. The system of Panini, In Language in India, vol 4:2. February 2004

Johansson Christer. (ed) 2007. Proceedings from the first Bergen Workshop on Anaphora Resolution (WAR I) , Cambridge Scholarly Publishing, UK

Kennedy Christopher and Branimir Boguraev. 1996. Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In Proceedings of the 16th International Conference on Computational Linguistics (COLING'96), 113-118, Denmark.

Khokhlova L.V. 1998. Some notes on reflexivization in Hindi and Russian, Moscow State University

Lappin, Shalom and Herbert Leass. 1994. An Algorithm for Pronominal Anaphora Resolution, In Computational Linguistics, 20, 4, 535-561.

Mitkov Ruslan, Boguraev Branimir, Lappin Shalom 2001. Introduction to the special issue on Coputational Anaphora Resolution. In proceedings of ACL 2001

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

Murthy K.N., Sobha L, Muthukumari B 2005. Pronominal resolution in Tamil using machine learning, In Proceedings of WAR I (Johansson ed.) Cambridge Scholar Publishing, 2007

Sigurd Bengt. 2005. Referents grammar and text. In proceedings of WAR I (Johansson ed.) Cambridge Scholar Publishing, 2007

Sobha L, Patnail B.N. 1998. An Algorithm for Pronoun and Reflexive Resolution in Malayalam. In Proceedings of the International Conference on Computational Linguistics, Speech and Document processing, C63-66.

Sobha L. Pralayankar Praveen. 2007. Anaphors in Sanskrit, presented at Referential Entity Resolution (RER), AU-KBC, Chennai, 2007