8.7 Handling Polysemous Particles in Multilingual Environment

Anil Thakur and Sanjukta Ghosh

Abstract—Particles across languages are tough to process. They exhibit complexity at various levels of processing. They have multiple functional roles based on different syntactic and contextual factors. Traditional grammars have inadequate account of them. In this paper, we propose a corpora-based comprehensive analysis of selected polysemous particles, particularly in a multilingual environment where languages are related. The paper also explores the mapping patterns of some of the particles across selected Indian languages and shows the significance of the nature of divergence among them and their NLP implications.

Index Terms—Particles, Disambiguation, NLP, Multilingual

I. INTRODUCTION

A number of indeclinable words in a natural language are found to be used in multiple grammatical roles (Thakur and Patnaik 2004, Sinha and Thakur 2005). Besides their primary grammatical roles, the other uses of such words are neither easily identifiable nor are they easily categorizable. We treat such "other" functions/uses of these words under their particles use. For instance, take the Hindi word *Or* 'and'. Its primary grammatical role is to conjoin two words, phrases or clauses/sentences, as in (1) below. However, it is also used in other contexts as in (2), where it does not occur in its typical conjunction role.

1. *raam **Or** siitaa donoN van gaye.*
   Ram and Sita both forest went
   'Both Ram and Sita went to forest.'

2. ***Or** ji, kyaa haal hE?*
   PRT PRT what state is
   'So how are you?'

These uses of *Or* in Hindi, as in (2) where it functions as a sentence starter (SS) particle, are the typical cases of the particle uses of indeclinable words in a natural language. The phenomenon is attested across languages. Among related Indian languages, some examples from Oriya can be illustrative (examples are from Thakur and Patnaik 2004). *je* in Oriya is primarily a complementizer that connects two clauses (the main clause and the subordinate clause), as in (3a). However, it can also occur in other atypical uses as in other examples in (3) below.

3.a. *mu jaane **je** se kaalii aasibe.*
   I know COMP he tomorrow come-FU
   'I know that he will come tomorrow.'

b. ***je** tume mo kathaare ete raagi jaucha!*
   PRT you my words so much angry getting
   'Oh you are getting so angry with what I'm saying!'

c. *mu tumaa kathaa kaahinki sunibi **je**!*
   I your words why listen-FU PRT
   'Why should I listen to you?'

In (3b), *je* occurs as a sentence starter particle whereas in (3c), it occurs as a pragmatic particle. In traditional dictionary, however, *Or* or *je* would be entered only in their primary role/grammatical category, that is, that of conjunction and complementizer respectively. As the discussion above clearly shows, the uses of such words in multiple senses need to be identified and separately categorized according to the different roles they occur with. Particularly, in the context of applied natural language processing, the task of identification and classification can play an important role. Since there are lots of commonalities across related Indian languages, it would be appropriate to examine the issue in a multilingual perspective. Our assumption is that the analysis of the nature of a particular particle in an Indian language can throw much light on the counterparts of those particles in other related Indian languages. With this assumption in mind, we examine the polysemous nature of a couple of particles and their distribution across Hindi, Bangla and Maithili. On the basis of our analysis, we propose that particles in a multilingual environment need to be examined with the help of large corpora and they need to be identified and categorized for their use in multiple roles. Some of the immediate advantages of such study include capturing the divergence patterns across related Indian languages (particularly for machine translation among them) and for tagging purpose for both monolingual and parallel corpora. The relevance of the study in the context of scientific and technical terminology can be seen in the light of the fact that there is much gap between availability and requirement of technical terms (particularly in Indian languages) to define and describe various concepts in many of the newly emerging branches of linguistics such as computational linguistics, neurolinguistics, cognitive linguistics, etc. The paper proposes to provide a ground for augmentation and standardization of scientific and technical terms, particularly in the area of linguistics.

Anil Thakur and Sanjukta Ghosh are with the Department of Linguistics, Banaras Hindu University, Varanasi (e-mail: anilt@bhu.ac.in, san_subh@yahoo.com)

II. SOME PARTICLES AND THEIR DISTRIBUTION

A. Some particle-uses of 'ki' in Hindi, Bangla and Maithili

ki: Alternative-Question Particle

In many Indian languages such as Bangla, Maithili, Bhojpuri, etc. *ki* is a question particle (its Hindi counterpart is kyaa). However, in Hindi too, besides in Bangla and Maithili, *ki* occurs in certain cases in the sense of an alternative-question particle. Examples in (4) show distribution of *ki* across Hindi, Bangla and Maithili[1].

4. a. mujhe nahiiN maalum ki raam aa-yaa (yaa) ki nahiiN?
   to-me not known COMP Ram came or PRT not
   b. aamii jaani na je raam aaSbe ki (aaSbe) naa?
   I know not COMP Ram come-FU PRT come not
   c. hamaraa nai bujhal achi ki raam Otaa ki nai (Otaa)?
   to-me not known COMP Ram come-FU PRT not come-FU
   'I do not know whether Ram comes or not?'

As we notice, the first use of *ki* in Hindi and Maithili is that of complementizer (COMP) (the Bangla counterpart is *je*). The second use of *ki* in Hindi and Maithili and the only *ki* in the Bangla sentence is used to mark alternative-question. The sentences of this type can be used either to ask question or to reply in a normal discourse. Since this *ki* is used to indicate question in the alternative possibility of answer in positive or negative, we can call this *ki* alternative-question particle.

ki: Purpose-Clause Particle

ki in Hindi and Maithili can also be used to indicate a purpose clause function. Normally purpose clause in Hindi and Maithili is marked by *taaki* 'so that' and *jeki* respectively. In Bangla, it is marked by *jate* 'so that'.

5. a. jaraa jor boliye ki ham bhii sune.
   please loud speak PRT we EMP listen
   b. kTu jore bolun jate aamraao suni.
   please loud speak PRT we-EMP listen
   c. kani jorsa baaju je/ki hamhusab suni.
   please loud speak PRT we-EMP listen
   'Please speak (a bit) louder so that we may also listen.'

[1] The (a), (b) and (c) examples onwards represent Hindi, Bangla and Maithili respectively.

We notice that both Hindi and Maithili use *ki* for this function, Bangla does not show this option, in which only *jaate* can be used.

ki: Correlative-Clause Particle

ki is also optionally used to mark correlative clause along with the correlative words (such as *jitanaa*, *jahaaN*, *jo*, *jab*, etc.) in both Hindi and Maithili. However, in Bangla this use of *ki* is not attested, as the examples in (6) illustrate. The optional use of *naaki* in Bangla is highly marked.

6. a. mEN utanaa nahiiN jaanataa jitanaa ki aap jaanate hEN.
   I that much not know as much PRT you know AUX
   b. aami OtoTaa jaanii naa jOtaTaa (?naaki) aapanii jaanen
   I that much know not as much PRT you know AUX
   c. ham otek nai jaanai chii jatek ki aahaaN jaanai chii.
   I that much not know aux as much PRT you know AUX
   'I do not know as much as you know.'

ki: Wish/Possibility-Clause Particle

ki is also used to introduce a clause to the main clause of wish-possibility type, as in (7-8). The Bangla counterpart of this *ki* is *je* and *jodi* in (7b) and (8b) respectively. Maithili[2] uses either *ki* or *je* or can also use both together in some cases.

7. a. kahiiN yEsaa na ho ki aap gir paRe.
   may be PRT you fall down
   b. jEno Emon na hoy je apnii poRe gelen.
   may be PRT you fall down
   c. kahiiN i na hoye je/ki ahaaN gir paRii.
   may be PRT you fall down
   '(Take care) may be, you fall down.'
8. a. acchaa hotaa ki ham pahale aaye hote.
   good happened PRT we earlier come aux
   b. bhaalo hoto jodi aamra aage aastam.
   good happened if we earlier come
   c. niik hoit je/ki ham pahine aayl rahitauN.
   good happened PRT we earlier come aux
   'It could have been nice if we had come earlier.'

In (8), Bangla uses only *jadi* 'if' whereas in Hindi and Maithili *ki* and *je/ki* respectively can be used to indicate this particular clause construction.

[2] The use of ki in Maithili may be because of the influence of Hindi.

ki: Temporal-Clause Particle

Both Hindi and Maithili have the temporal adverbial use of *ki*. Bangla, however, does not have such use. In Bangla, the identical construction is expressed by the exclusive focus suffix, the counterpart of hi in Hindi. The similar use of hi is also attested in Hindi.

9. a. raam jEse hii ghar pahuNcaa ki baarish band ho gayii.
   Ram as soon as home reached PRT rain stopped be went
   b. raam baaRii pouchono matroi/pouchotei brisTi bOndho hoye gElo.
   Ram home reaching as soon as rain stop be went
   c. raam jahinaa ghOr pahuNcalaa ki brisTi band bhO gelai.
   Ram as soon as home reached PRT rain stop be went
   'As soon as Ram reached home, it stopped raining.'

The particle-uses of *ki* shown in these examples are only illustrative of the point discussed in this paper. A detailed study of the multiple functions of *ki* along with their mapping patterns (from the point of view of Hindi-English machine translation) is given in Sinha and Thakur (2005). Besides *ki*, many other words can be shown to have this property of occurring in multiple functional roles. We can take *kyaa* in Hindi and its Bangla and Maithili counterpart *ki* to illustrate the point.

B. Some particle-uses of 'kyaa/ki' in Hindi, Bangla and Maithili

kyaa/ki is primarily an interrogative pronoun as well as a question particle (for yes-no question) (a detailed study on the disambiguation of *kyaa* in Hindi is given in Sinha and Thakur forthcoming) in these languages. However, *kyaa/ki* also occurs in other functional roles in these languages, a few of which can be shown in this section.

kyaa/ki: Emphatic Assertion Particle

kyaa and *ki* are used to indicate emphatic assertion in Hindi and Maithili respectively. However, in Bangla, this function is marked by a topic particle to as shown in (10). Hindi and Maithili also use topic particle *to* and *ta* respectively for topic marking.

10. a. yah kyaa paRii hE tumahaarii kitaab !
    this PRT lie.PRF be.PR your book
    b. ei to poRe roeche tomar boi !
    this PRT lie PR.PRF your book
    c. i ki (raakhal) chau tohar bOhii.
    this PRT kept is your book
    'Your book is lying very much here.'

kyaa/ki: Emphatic Negation Particle

kyaa is also used to indicate emphatic negation and its Bangla and Maithili counterparts *ki* is used for the similar function in these languages, as in (11) below.

11. a. vah tumhaaraa kyaa bigaar legaa.
    he your PRT harm do.FU
    b. o tomar ki khoti korbe. (B)
    s/he your PRT harm do.FU
    c. se tohar ki bigaariletau. (M)
    s/he your PRT harm.FU
    'He cannot do any harm to you.'

kyaa/ki: Temporal Adverbial Particle

Another use of *kyaa* in Hindi and its Bangla and Maithili counterpart ki is to mark temporal adverbial, as in (12).

12. a. vah dikhaa kyaa ki sab uskii or lapak paRe.
    he seen PRT that all his towards jump fell
    b. oke dekhlo ki SObai or opor jhaNpiye poRlo.
    her/him saw PRT all her/his upon jump.CP fell
    c. o dekhi ki paRal ki sOb okare dish lapaki paRal.
    he seen PRT happened that all his direction jump fell
    'The moment he was seen, all jumped towards him.'

However, we may also notice that in Bangla, the syntactic environment of this use of *ki* is different from that of *kyaa* in Hindi and *ki* in Maithili. In the discussion above, we have listed a couple of particles and some of their multiple uses in Hindi, Bangla and Maithili. This has been done with a view to showing the state of indeterminacy about the different functions of such words and their appropriate categorization.

III. IDENTIFICATION AND CATEGORIZATION

As we have seen from the discussion of only two particles, Hindi and its Bangla and Maithili counterparts have much in common with respect to their

distributional behavior. However, we have also noticed some of the differences that they have. Both their common features and those on which they differ can be used to identify and categorize the behavior of the particles in general. Secondly, the identification of their commonalities and differences is useful for building comparative lexical resources on particles across these languages. As we can notice, the identification of the different roles of such particles is dependent on both their syntactic and semantic analysis such as the position of their occurrence in the sentence, the nature/type of the sentence, the occurrence of some other element in the sentence, etc (Sinha and Thakur 2005). Thus, we notice significant commonalities across languages with respect to the identification strategies, too. On the basis of these observations, we can propose that a study for identification and categorization of the polysemous words/particles can be more fruitful in multilingual environment

IV. DIVERGENCE AND NLP APPLICATIONS

It is known that a study of structural similarities as well as divergence patterns among the related Indian languages will go a long way in obtaining mapping patterns across these languages. For instance, for the purpose of machine translation across Indian languages a study of this kind can be quite indispensable. Another advantage of such study can be seen with respect to the enrichment of parallel lexical resources that can be useful for various natural language applications. Exact identification and categorization of particles, for example, will lead to appropriate tagging. For various NLP applications in Indian languages (e.g., Hindi, Bangla, Maithili), a contrastive analysis of the languages is important.

REFERENCES

[1] Patnaik, B. N. (1979) "Synatx and Semantics of the Indeclinable *boli* in Oriya." *Language Forum 5.1*.

[2] Sinha, R.M.K. and Thakur, Anil (2005). "Handling *ki* in Hindi for Hindi-English MT". *Proceedings of MT SummitX*, Thailand.

[3] Sinha, R.M.K. and Thakur, Anil (forthcoming). "Disambiguation of *kyaa* in Hindi for Hindi to English Machine Translation". *Indian Linguistics*.

[4] Thakur, Anil and B.N. Patnaik (2004). "Indeclinables and Natural Language Processing". *Proceeding of the International Symposium on MT, NLP and Translation Support Systems (iSTRANS)*. New Delhi: Tata McGraw-Hill Publishing Co.

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.