# Disambiguation of *karmaṇī* (k2) of Hindi in Hindi-Telugu Machine Translation

*Christopher M., Uma Maheshwar Rao G.*

Center for Applied Linguistics and Translation Studies
University of Hyderabad
efthachris@gmail.com, guraohyd@yahoo.com

## Abstract:

Rule based language modeling of grammar is one of the important components of Natural Language Processing (NLP) in general and Machine Translation (MT) in particular. For any MT system, the case mapping and its disambiguation is a challenging task. In this paper, an attempt is made to disambiguate *karmaṇī* (K2) of Hindi and mapping the case to its corresponding Telugu equivalent, and its implementation in a Machine Translation System. Our analysis is based on the assumption that semantic information associated with the Nouns (N) and Verbs (V) plays a major role in disambiguating the multi-mapping of case of the *karmaṇī-K2* of Hindi to Telugu. This paper also discusses the computational implementation of the above mentioned issue and the process of testing it in a Hindi-Telugu MT system.

**Keywords:** - Case Marker, *Ka:raka Role*, Panini, Thematic Role, NLP, MT, *karmaṇī,* Ontology, Valency.

## 0. Introduction:

Indian Languages have a relatively free word order. Many of the constituents of a simple sentence can occur in any order without affecting the gross meaning of the sentence; what effected is perhaps the emphasis etc. For instance, Noun Phrases (NPs) in a sentence can come in any order without affecting the *ka:raka* relationship between the Verb Phrase(VP) and Noun Phrase(NP). Since position or order of occurrence of a NP does not contain information about the *ka:raka* or Theta roles in a simple sentence. A question can be asked regarding what carries this information. The answer seems to be that Post-Positional (PP's) Markers or Case Markers (CM's) after nouns in Indian Languages play a key role in specifying semantic Relationships. Postposition and suface case endings of nouns can collectively called as *vibhaktis* of Nouns'. *Vibhaktis* are very crucial in determining the semantic role of the NPs with the VPs of the sentence. But, things are not always stright forward and the following needs to be accounted for:

**a) Many-to-One:** A different *vibhaktis* can be used for the same semantic relation ***(Agent)*** with a given verb with different Tense.

Ex: 1) rAma ***(∅)$_{nom.}$*** Pala ko KawA hE.
2) rAma ***ne$_{erg.}$*** Pala KayA
3) rAma ***ko$_{dat.}$*** Pala KAnA pada
4) rAma ***se$_{inst.}$*** Pala nahiM KayA gaya.

**b) One-to-Many:** The same vibhakti i.e ***ne*** can be used with the same verb for two different semantic relations (Agent, Instrument).

Ex: 1) mohan ***ne$_{erg.}$*** wAlA kola.
2) cAbI ***ne$_{inst.}$*** wAlA kola.

In a Machine Translation system, miss-match between *vibhaktis* and *ka:raka* roles leads to ambiguity. These ambiguity brings down the performance of the system and lead to poor

accuracy. So, to gain high accuracy and fluency of the translated text, disambiguation of these case miss-matches has become crucial.

This paper tries to address the above mentioned problem using Rule Based Method. Using rules to disambiguate *karmaṇī k2 of* Hindi into Telugu, in a Computational perspective is easy to handle and computationally efficient. Semantic (Ontological) features of nouns and verb with the help of verb valence are crucial in disambiguating the *karmaṇī k2* of Hindi while mapping it between dative –accusative of Telugu. This Algorithm has been used in Hindi-Telugu MT system that is being developed at CALTS[1] under IL-ILMT[2] project (www.tdil-dc.in).

The present paper is divided into five sections. First section deals with the case markers of both languages viz. Hindi and Telugu with their *Thematic and ka:raka roles*. Section two tries to explane about ***karmaṇī*** *(k2)* with the help of *patanjali's* comments and provides suitable examples in Telugu and Hindi. In third section, a detailed description about the Computational Implementation of the karmaṇī disambiguation Engine is given. In Fourth section the system is Evaluated and Results are listed. The Fifth and final section dicusses the conclusions drawn in the paper.

## 1. Case Markers, *ka:rakas* and Thematic Roles:

The notation of *ka:raka* relations is central to *Paninan* framework. The *ka:raka* relations are syntactico-semantic relations between the verb and its arguments in a sentence. In other words, a *ka:raka* is a person or an object which does something or which participates in carrying out an action in some way or other (*karoti:ti ka:rakam*). Deep or underlying relationship that holds between a noun and a verb is displayed by *ka:raka* relationships. Each *ka:raka* normally represents a single semantic concept but in a few cases, which are well defined may represent more than one semantic concept.

*ka:rakas* capture a certain level of semantics. The approach uses case markers (*vibhakti* information) for mapping the relation between the verbs and their arguments. The six basic *ka:rakas* are : (note that the English translations are only approximations and do not fully capture the concepts below. However, it must be noted that although one can roughly map the last four *ka:rakas* to their thematic role counterparts, *karmaṇī* and *karwa* are different from theme and agent (though they might map with them sometimes). The reason for this divergence in the two notations, *ka:raka* and thematic role, is due to the difference in what they convey. A list of case markers of both Hindi and Telugu with their *ka:raka* roles, Thematic roles in a surface level are given.

| S. No | **Cases (Western Name)** | ***Ka:rakas* (Paninian)** | **Thematic Roles** | **Case Markers** | | **Prototypical Definition** |
|---|---|---|---|---|---|---|
| | | | | **Hindi** | **Telugu** | |
| 1 | Nominative | karwa (k1) | Agent, Experiencer, Force | ∅ | ∅ | The Independent one (Agent) and Volitional |
| 2 | Ergative | | | ***ne*** | | |
| 3 | Accusative | karmaṇī (k2) | Theme, Patient, Content, Result, Goal | ***ko*** | ***ni*** | The thing most desired by the Agent (patient, theme) |
| 4 | Dative | saMpraxana (k4) | Beneficiary | ***ko*** | ***ki*** | The item in view through the karma (goal) |
| 5 | Instrumental | karaṇa (k3) | Instrument | ***se*** | ***wo*** | The most effective means |

[1]Center for Applied Linguistics and Translation Studies (CALTS), School of Humanities, University of Hyderabad.

[2] Indian Language to Indian Language Machine Translation Project funded by MCIT, DIT, TDIL, Govt. of India

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | (Instrument) |
| 6 | Ablative | apa:xa:na (k5) | Source | ***Se*** | ***nuMdi*** | The fixed point from which something recedes. |
| 7 | Genitive | sasti (r6) | Possessive | ***kA/ke/kI*** | ***yoVkka*** | Adnominal |
| 8 | Locative | adhikaraṇa (k7) | Time, Place | ***meM /para*** | ***lo,na, mi:da*** | Locus, Location of the incident |

Table-1

## 2. *karmaṇī* (k2) in Hindi and Telugu:

*karmaṇī* is an object/patient of the verb, denoted as k2. *karmaṇī* or k2 can also be marked on two nouns or NP's with respective to the verb which expresses a change of state. The destination or goal of a motion verb may also be marked as k2. Comments of *Patanjali* on *karmaṇī* can be described as:

1) *kartur i:pasitatamam karma*

That which is most desired or intended by the agent through an act is *kaman:*

Ex: **Hi.** *usne kitAba* ***ko*** *becA*
**Gloss:** he-erg book-acc. sell
***Te.*** *vAdu pustakAn****ni*** *ammAdu*
**En.** He sold the book

2) *tatha: yuktamc a:ni:psitam*

That which is not most desired or intended by the agent but which still is connected with the action like the one which is most desired or intended is also karmaṇī.

Ex: **Hi.** *rAswemeM usne sApa* ***ko*** *deKa*
**Gloss:** way to he-erg snake-acc. see-past-sg
***Te.*** *xArilo awadu pAmu****ni*** *cusAdu*
**En.** In his/her way he/she saw a snake.

3) *akathitam ca*

That ka:raka which is not considered to be one of the above *ka:raka* is also *karmaṇī*, only when associated with an intransitive verb, denoting a place, time condition and distance to be covered are considered to be *karmaṇī*.

Ex: **Hi.** *kOna jane kal* ***ko*** *kay hoga*
**Gloss:** who know tomorrow –dat. what happen
***Te.*** *reVpati****ki*** *emi jarugu wuMdo evariki weVlusu.*
**En.** Who knows what happens tomorrow.

4) *gatibuddhi pratyavasa:na:rtha sabdakarma: karmak: na:mani karta: sa nau.*

If the verb having the sense of "motion", knowledge or information and consumption verbs have some literary work for their object and of intransitive verbs, that which is the agent in their non-causative state will become *karmaṇī* in their causative state.

Ex: **Hi.** un***ko*** ElAna karo
**Gloss:** Them-dat. alert make
**Te.** valYla***ki*** jagratta ceVppu
**En.** Warn them.

*5) harkror aiya tarasya:m*

The agent of the verb *har* (to take) and *kror* (to make) optionally become *karmaṇī* when the verb takes the causative suffix.

Ex: **Hi.** XoBi ***se*** kapade XulavAyA
**Gloss:** washer-man-asso. clothes wash-caus
**Te.** sAkalivAdi***cewa/wo*** battalu vuwikiMcAnu
**En.** I made washermen to wash the cloths.

## 3. Implementation in MT:

This section describes the implementation of case disambiguation from Hindi to Telugu with special reference of karmaṇī. To understand this implementation we first go through the steps of Machine Translation: sentences are divided into phrases and local words are grouped with respective to the heads of phrases. Simple parser identifies dependency (*ka:raka*) relations. Later WSD engine disambiguates the tokens and lexical substitution engine transfers lexical and functional items into the target language. Even the functional elements have to be disambiguated and correct mapping should be done to gain high accuracy, fig-1 below gives us a detail description about the procedures of implementation.

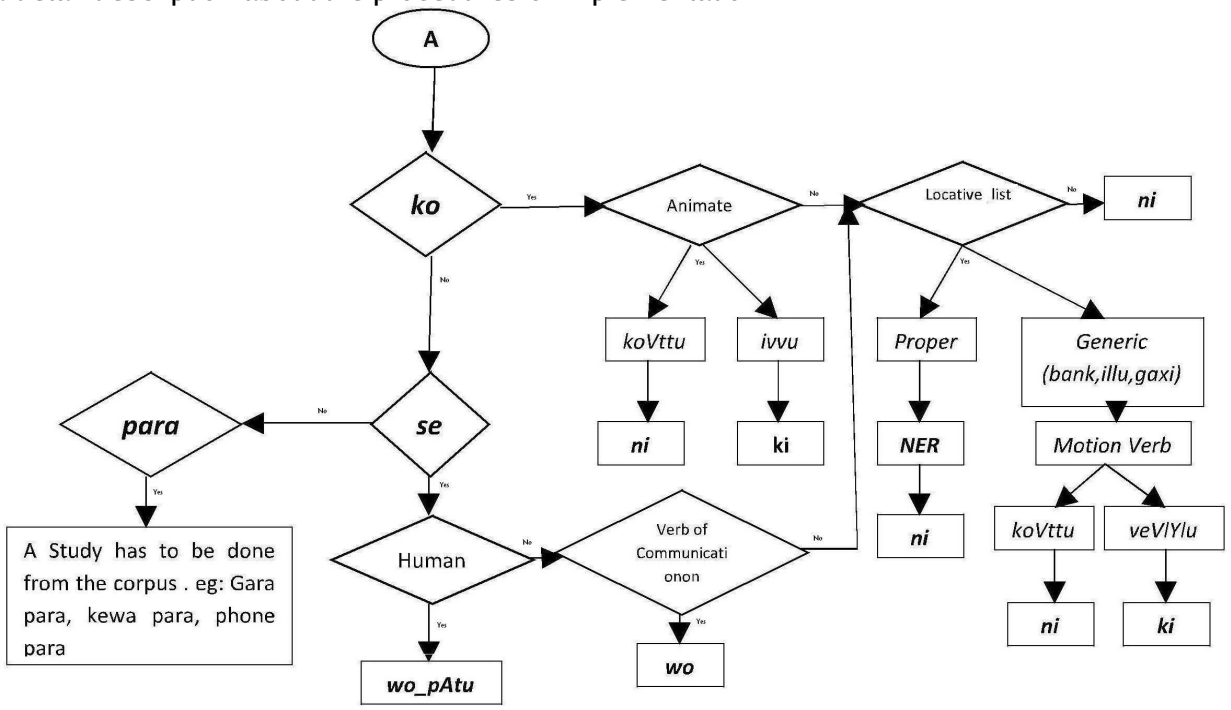


Fig-1

## 4. Evaluation and Results:

We demonstrate a system which performs case disambiguation from Hindi to Telugu. Evaluation of the system is important to validate our approach. We performed a user based evaluation. The system output were shown to the human evaluators and they were asked to rate the output based on correct replacement of the case markers. Depending upon their feedback the systems performance is measured.

For the purpose of evaluation, we took 738 sentences from 3 million word Hindi Corpus.

Of the 738 sentences 380 were simple constructions, 160 are coordinate and subordinate constructions, 100 are complex constructions and 92 are relative constructions.

For testing, dependency relations are marked for all these sentences, in which 1053 were marked for karmaṇī as shown in the Table-2. We then run these sentences through case disambiguation engine. The correct replacements were about 81.29%.

| S.No | Sentence Type | No. of Sent | No.of K2 | Correct Sub. | % correct sub |
|---|---|---|---|---|---|
| 1 | Simple | 380 | 440 | 406 | 92.27 |
| 2 | Coordinate & Subordinate | 160 | 284 | 258 | 90.84 |
| 3 | Complex | 100 | 177 | 109 | 61.58 |
| 4 | Relative | 92 | 152 | 83 | 54.6 |
| Total | | 732 | 1053 | 856 | 81.29 |

Table-2

## 4.1. Error Analysis:

The error analysis for the incorrect replacement of the karmaṇī are persented in Table-3. A large percentage of errors were due to the wrong POS-Tagging and Chunking, even head identification for each phrase contributes for wrong replacement of karmaṇī. Lexical disambiguator or WSD plays a vital role, if the verb is not disambiguated before substituting it into the target language (Telugu). Wrong verb mapping leads to wrong valancy which multiplies with wrong Ontology and gives us wrong substitution of cases.

| S.No | Module Name | % of Error |
|---|---|---|
| 1. | POS-Tagger | 19 |
| 2. | Chunker+Head Identification | 26 |
| 3. | Lexical Substitution + WSD | 49 |
| 4. | Valency | 6 |

Table-3

## 5. Conclusion:

Here we give examples based on rules which are perfectly working for the MT system from Hindi to Telugu, being developed by CALTS, HCU under IL-ILMT project www.tdil-dc.in. The system is still in progess and the rules discussed above are representation and need future examination to make them more precise and also we need to look on other cases divergences.

## Reference:

1.Akshar Bharati, Vineet Chaitanya, Rajeev Sangal. 2001. *"Natural Language Processing: A Paninian Perspective"*. Premtice-Hall of India, New Delhi.

2.Kachru, Yamuna. 1980. *Aspects of Hindi Syntax*. Munshi lal Manohar lal, Delhi.

3.Mohanan, Tara. 1994. *Agrument Strucutre in Hindi*. CLSI publication, standford, California.

4.Rao, Rama C. 1976. *Markedness in Case*. "D ravidian Case system" ed: S. Agsthialingam ( et.al ). Annamalai University, India.

5.Subrahmanyam, P.S. 1976. *Ka:rakas and Case Markers*. "D ravidian Case system" ed: S. Agsthialingam ( et.al ). Annamalai University, India.