Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 502 – 510

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES-2016

A pāniniān framework for analyzing case marker errors in English-Urdu machine translation

Sharmin Muzaffar[a], Pitambar Behera[b*] & Girish Nath Jha[ab]

[a]*Department of Linguistics , Aligarh Muslim University, Aligarh, India*
[b*, ab]*Centre for Linguistics, Jawaharlal Nehru University, New Delhi, India*
{sharmin.muzaffar,pitambarbehera2, girishjha}@gmail.com

Abstract

Panini's Kāraka Theory is solely based on the syntactico-semantic approach to understanding a natural language which takes into consideration the arguments of the verbs. It provides a framework for exhibiting the syntactic relations among constituents in terms of modifier-modified and semantic relations with respect to Kāraka-Vibhakṯi (semantic role and postposition).
In this paper, it has been argued that Pāniniān Dependency Framework can be considered to deal with the MT errors with special reference to case. Firstly, a corpus of approximately 500 English sentences as input have been provided to Google and Bing online MT platforms. Thereafter, all the output sentences in Urdu have been collated in bulk. Thirdly, all the sentences have been evaluated and errors pertaining to case have been categorized based on the Gold Standard. Finally, Pāniniān dependency framework has been proposed for addressing the case-related errors for Indian languages.

Peer-review under responsibility of KES International

Keywords: Pāniniān Dependency Framework; Kāraka-Vibhakṯi; English-Urdu MT, Kāraka Theory.

* Pitambar Behera. Tel.:+91-9971691598; E-mail address:pitambarbehera2@gmail.com

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.117

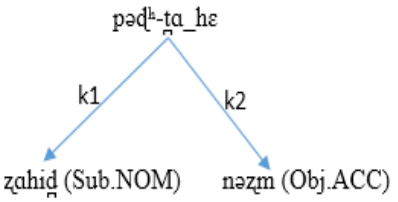
1. Overview

Indian languages like Urdu, Hindi, Telugu and many others are morphologically rich languages [19] and have relatively flexible word-order in comparison to European languages like English, German and so on. Linguistically, Urdu and English have divergent features [18]. The reason is that Urdu and English belong to different language families [18], have divergent grammatical and semantic structures [19]; ILs have free word-order [2, 3] and above all they have different cultural backgrounds. One of the divergences is that English has prepositions in prepositional phrases while Indian languages Urdu have postpositions in postpositional phrases.

a) ʐ̪ahiḏ̪ nəʐ̪m pəɖʰ-ʈ̪a hɛ (SOV)
3MSG.NOM poem-3FPL read-3MSG.IMPFV PRS
"Zahid reads poem."
b) nəʐ̪m ʐ̪ahiḏ̪ pəɖʰ-ʈ̪a hɛ (OSV)
c) nəʐ̪m pəɖʰ-ʈ̪a hɛ ʐ̪ahiḏ̪ (OVS)
d) ʐ̪ahiḏ̪ pəɖʰ-ʈ̪a hɛ nəʐ̪m (SVO)

Out of these above-instantiated four possible word-orders, the first one is the unmarked whereas the rest of the following are marked and acceptable in Urdu. The first instance shows the enriched morpho-syntactic information (PNG and TAM) encoded in different grammatical categories of the sentence.

The Dependency tree which accounts for all of the instances is as follows:



In the above dependency tree, the agent of the action /pəɖʰ-ʈ̪a_hɛ/ is /ʐ̪ahiḏ̪/ and the patient is /nəʐ̪m/. The tree accounts for all examples as well which allows a scrambled word order.

Furthermore, some of the acceptable, grammatical and semantically well-formed English sentences translate into Urdu inappropriately. For example,

a) The shop sells well.

b) *ɖ̪ʊkɑn əccʰe se becʈ̪a hɛ.

The appropriate English in the above sentence (a) maps into Urdu counterpart inappropriately because the latter does not allow such semantic information. For the sentence to be semantically well-formed, the agent has to pass the subjecthood test [16, 17] and needs to have the global semantic features [+animate, +human] encoded in Urdu. Therefore, it is pertinent to experiment with the Pāniniān dependency framework which represents the kāraka relation which suggests the relationship of the nouns with the verb.

2. Pāniniān grammar (PG)

The PG [13, 14, 16, 4, 7, 8, 21, 20, 12] considers language as a medium of communication and the "information as central to the study of language". The speaker as an encoder expresses his ideas through language string* and the hearer decodes the information encoded in the communication to understand the meaning. PG solely deals with the

* 'String' refers to word/phrase/sentence/paragraph etc.

process of communication and provides a theoretical framework to model and extract the semantic information encoded in the process.

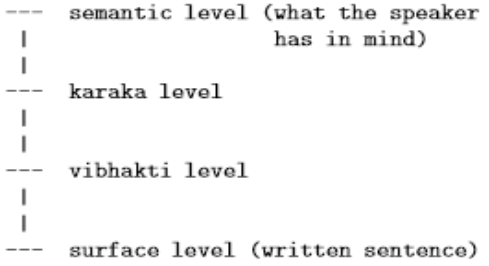


Fig. 1. Levels in the pāniniān model (Adapted from Bharti et al., 1996)

2.1. The kāraka theory

Trask has defined case as 'one of the forms which a noun or pronoun may assume in order to represent its grammatical and semantic relation to the rest of the sentence' [22]. There are different criteria for deciding the types: morphological, structural and semantic. Broadly speaking, cases are divided as direct and oblique. While the former covers only the nominative case the rest (accusative, dative, instrumental, ablative and locative) are covered by the latter. Case is realized in the form of postpositions in Indo-Aryan languages including Urdu; when they take nouns grammatically from phrases. Thus, they are known as postpositional phrases. These sorts of phrases consist of noun phrase followed by a postposition.

The PG framework has two major levels: the kāraka and vibhaḵ̱ti. The former suggests the relation between the verb and the other nouns in the sentence whereas the latter denotes to the local word groups based on case endings, prepositions or postposition markers. The kāraka relation is the syntactico-semantic relation close to the thematic relation which is reflected in the surface form. Case markers for nouns are generally the case endings and postpositions while for verbs are the TAM features encoded in the auxiliaries. There are six kāraka relations (see table. 1) along with their corresponding case markers: karṯ̱ā (agent), karma (patient), karana (instrument), sampradāna (beneficiary), apādāna (ablation) and adhikarana (locus).

Table 1. Kāraka and vibhaḵ̱ti table

Kāraka (Case Relations)	Vibhaḵ̱ti (Case Markers)
Karṯ̱ā (agent)	φ/ne
Karma (patient)	ko/φ
Karana (instrument)	se/ḏʊɑrɑ
Sampradāna (beneficiary)	ko
Apādāna (source)	se
Adhikarana (location)	mɛ̃/pər

2.1.1. The identification of kārakas

The mapping of the kāraka and vibhaḵ̱ti solely depends upon the two important structures: default kāraka chart and kāraka chart transformation [4]. The former specifies the case markers permissible by the specific kāraka relations for the nouns depending upon the TAM features of the verbs. One needs to have the knowledge about which kārakas a given verb can take to identify the kārakas that correspond to an activity.

2.1.1.1. Intransitive verbs

The intransitive verbs need to have a karṯ̱ā (agent) mandatorily while the karma is absent and other kārakas namely, instrument, location, ablation, beneficiary are optional components. Thus, in the example /ʐɑhiḏ̱ ḏ̱ɔʊḏ̱tɑ ṯ̱ʰɑ/, the

verb is in the imperfective aspect represented by /t̪ɑ/ and /t̪ʰɑ/ refers to the features [3MSG.PAST] [18]. The default kāraka chart for this above-instantiated sentence is as follows:

Table 2. Default kāraka chart for intransitives

Verb: ɖəɖnɑ TAM: t̪ɑ_t̪ʰɑ		
Kāraka	Vibhakt̪i	Optionality
Kart̪ā (agent)	φ	mandatory
Karma (patient)	ko or φ	optional
Adhikarana (locus)	mɛ̃/pər	optional

2.1.1.2. Transitive verbs

With regard to the transitive verbs, it can be stated that they ought to have kart̪ā (agent) and karma (patient) mandatorily while other Kārakas have to be optional components. In the example /ʐɑhiɖ kəhɑni pəɖʰ-t̪ɑ hɛ/, the verb /pəɖʰnɑ/ is a transitive verb by default and is represented as follows.
Verb: pəɖʰnɑ TAM: pəɖʰ-t̪ɑ_hɛ

Table 3. Default kāraka chart for transitive verbs

Kāraka	Vibhakt̪i	Optionality
Kart̪ā (agent)	φ	mandatory
Karma (patient)	φ	mandatory
Karana (instrument)	se/ɖʊɑrɑ	optional
Adhikarana (locus)	mɛ̃/pər	optional

In the instance /ʐɑhiɖ-ne kəhɑni pəɖʰi/, the agreement is licensed in the verb by the object because of the perfectivity and transitivity[†]. This information is represented by the default Kāraka chart as follows:

Table 4. Default kāraka chart for transitive verbs with /ne/

Verb: pəɖʰnɑ TAM: pəɖʰ-i		
Kāraka	Vibhakt̪i	Optionality
Kart̪ā (agent)	ne	mandatory
Karma (patient)	ko or φ	mandatory
Karana (instrument)	se/ɖʊɑrɑ	optional
Adhikarana (locus)	mɛ̃/pər	optional

The transformation from the nominative to ergative and nominative to dative-subject can be represented by the Kāraka chart transformation as follows without preparing any further default Kāraka chart. For the TAM features /i,ɑ,jɑ,ji/ that suggest the perfectivity and transitivity of the aspect /ne/[‡] vibhakt̪i marker is applicable. Dative subject with the infinitive endings /nɑ-pəɖɑ/ will have the vibhakt̪i marker /ko/.

Table 5. Transformation rules

TAM Features	Rules

[†] *Ibid.*
[‡] *Ibid.*

i/ɑ/jɑ/ji	Vibhakt̪i (kart̪ā) = ne
nɑ-pəɖɑ	Vibhakt̪i (kart̪ā) = ko

2.1.1.3. Di-transitive verbs:

Di-transitive verbs in the perfective aspect will have the following kāraka chart where there are three arguments of the verb and all are mandatory: kart̪ā, karma and sampradāna. In the sentence 2 /ʃʊmɛlɑ ne rəhim-kɒ ek kəmiz̪ d̪ed̪i/ (see table. 6), there are only three arguments of the verb i.e. Shumaila, Rahim and the shirt. These three are mandatory whereas the others instrument and location are optional elements.

Table 6. Default kāraka chart for di-transitive verbs

Verb /d̪enɑ/ TAM d̪e_d̪i

Kāraka	Vibhakt̪i	Optionality
Kart̪ā (agent)	ne	mandatory
Karma (patient)	ko or φ	mandatory
Karana (instrument)	se/d̪ʊɑrɑ	optional
Sampradāna (beneficiary)	ko	mandatory
Adhikarana (locus)	mẽ/pər	optional

3. PG dependency analysis of case markers errors

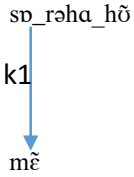
Kāraka relations suggest the utmost amount of semantic information which can be extracted from a language neither taking recourse to the extra-linguistic features nor the contextual linguistic knowledge which readily is available at hand. This section demonstrates different kinds of kāraka relations, the errors committed by the MT platforms (Google and Bing) pertaining to the case markers, identification and resolution through the PG dependency relation. For the annotation of dependency relation in syntactico-semantic parsing, the IIIT Hyderabad annotation convention [3, 5, 6, 9, 21] (see table 7 below) has been adhered.

Table 7. Annotation labels for PG dependency parsing

Annotation labels	Description
k1	kart̪ā (similar to agent/doer)
k2	karma (similar to patient/theme)
k3	instrument
k4	beneficiary
k5	source
k7t	temporal location
k7p	spatial location
k1s	noun complement
k2p	destination
pk1, mk1, jk1	Causer, mediator-causer, causee
rh	cause
rt	purpose
rsp	duration
adv	adverb (manner)
pof	Part-of (complex predicates)
ccof	conjunction
fragof	fragment-of

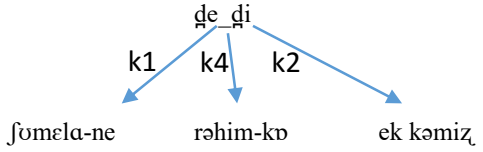
3.1. Karṭā kāraka or nominative case

In the karṭā-kāraka relation, the karṭā (agent) is the most independent participant in the action and there is no overt case marker for this kāraka. In other words, the āsraya (the locus) of activity resides in the karṭā and thus there is one semantic role assignment and the verb is intransitive.
In the sentence 1 (see table. 8), there is only one argument of the verb i.e. the pronominal I. From the Verb: sɒnɑ TAM: sɒ_rəhɑ_hɔ̃, it can be predictable that the verb 'sleeping' needs only one argument and hence it is intransitive. The default kāraka chart for this kāraka relation can be related to the chart made above (table. 2). As outlined above, the karṭā is mandatory while others are optional. (I am sleeping)



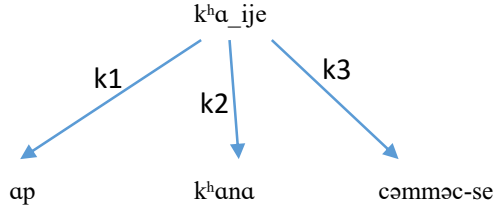
3.2. Karma kāraka or accusative case

When the asraya of the result is different from karṭā, then it is called karma. A verb which has asraya of activity and result can be different is called a (sakarmaka) transitive verb.
In the sentence 2 (see table. 7), there are only three arguments of the verb i.e. Shumaila, Rahim and the shirt. In both Google and Bing platforms, the translation outputs are wrong as the ergative and dative markers are missing. To predict the case markers, one has to analyse the verb and TAM features. The verb is /ɖenɑ/ and the TAM is /ɖeɖi/ which can be applied to intuitively predict that the verb takes more than one argument definitely. The kāraka chart for this sentence can be pertained to the chart for transitive verb where the karṭā and karma are mandatory but the others are optional. Therefore, the karma kāraka will get the role of the patient of the action which is the direct object i.e. the shirt. (Shumela has given Rahim a shirt.)



3.3. Karana kāraka or instrumental case

This kāraka is otherwise known as instrumental case. With the vyapara (activity) of the karma, Pala (result) is immediately achieved. In the sentence 3 (see table. 8), verb: kʰɑnɑ will have TAM features: kʰɑ_ije. The arguments of the verb such as karṭā, karma and karana will be mandatory whereas others are optional. Since the type of sentence is imperative, it is obvious that karṭā (agent) is the second person pronominal. (Please take food with the spoon.)

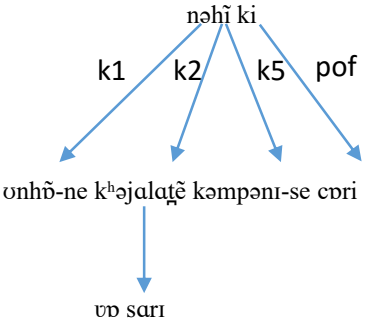


3.4. Sampradāna kāraka or dative case

Sampradāna kāraka is the indirect object which is the beneficiary of the action. In the sentence 2 (see table. 8), there are only three arguments of the verb i.e. Shumaila, Rahim and the shirt out of which Rahim gets the role of the indirect object or dative case. The verb is /ɖenɑ/ and the TAM is /ɖeɖi/ which refers to the transitivity of the verb and it takes more than one argument. The kāraka chart for this relation can be related to the chart of di-transitive verb (see table.5). Thus, the karṭā, karma and sampradāna are mandatory and the others are optional.

3.5. Apādāna kāraka or ablative case

Apādāna kāraka refers to the ablation or separation of the participant in an action. In the sentence 5 (see table. 8), the 'company' gets the apādāna kāraka which requires a karṭā, karma and an apādāna mandatorily while the others are optional. Both the platforms have correctly translated the English sentence. (He did not steal all those ideas from the company.)



3.6. Adhikarana kāraka or locative case

It refers to the locus or the temporal or spatial location of Karṭā or karma. As exemplified in the instance 4 (see table. 8), both the platforms get the One of the TAM features i.e. tense wrong. The verb will take two arguments the karṭā and another one under the adhikarana kāraka phrase.

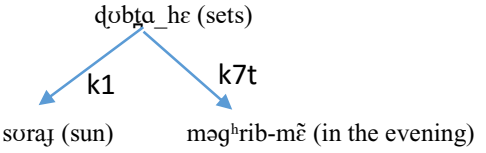


Table 8. Exemplary sentences for case markers errors on Google and Bing

Sl. no	Cases	Google Urdu output	Bing Urdu output	English input	Gold
1	NOM	mei so raha hu	mei so rahi hu	I am sleeping.	Main so raha/rahi hun
2	ACC/DAT	shumaila rahim ek qamiz di	shumaila rahim ek qamiz di	Shumaila gave Rahim a shirt.	shumaila ne rahim ko ek kemiz de di
3	INSTR	chamach se kha lo	chamache ke saath khana	Eat with spoon.	cammac se khana khaiye
4	LOC	suraj maghrib mein dubta	suraj ke gharoob ke maghrib mein hai	The sun sets in the evening.	suraj maghrib mein dubta hai
5	ABL	unhon ne kampanii se in qhayaalaat corii nahin ki	unhon ne kampanii se wo sab qhayaalaat corii nahin ki	He didn't steal those ideas from the company.	unhon ne kampanii se wo sarii qhayaalaatein corii nahin ki
6	GEN	uske shohar ko hamesha unki sehat ke bare mein shikayat hai	uske shohar ko hamesha unki sehat ke bare mei shikayat hai	Her husband is always complaining about his health.	uske shohar hamesha apni sehat ke bare mein shikayat karte hain
7	ERG	main ne tawaja nahin di	mujhe notice nahin kiya hai	I didn't notice.	main ne tawaja nahin di

4. Proposed algorithm and architecture of the kāraka parser

This section proposes a set of heuristic rules and the processes for making an algorithmic model in order to get parsing output of a sentence with relevant kāraka information. Given an input sentence, the default kāraka chart and the transformation rules, the parser algorithm will approach for kāraka parsing in the following manner.
Firstly, the parser analyses the input text breaking into morphological units by the morphological analyser. Secondly, all the words are grouped based on their respective heads by the Local Word Grouping (LWG). For instance, prepositions, adjectives and other noun modifiers are grouped as a chunk with noun as the head. Similarly, the all auxiliaries and adverbs are chunked under the verbal head. Thirdly, the parser divides all the words into two broader categories: demand words (verbs with TAM features) and source words (nouns with case markers). Fourthly, it applies the default kāraka chart and makes use of transformation rules if needed. Fifthly, it parses the input sentence into kāraka output if three following conditions are fulfilled.

- If every mandatory kāraka role is assigned to only one word in the output under processing.
- If every optional kāraka role is assigned to only one word
- If every word has the only and single kāraka role

If these above conditions are fulfilled and every word has only kāraka assignment, then the parsed output is the solution. If these conditions are not fulfilled or any one of the kārakas does not get an assignment or any of the kārakas gets more than assignment, then the parser will produce all the outputs.

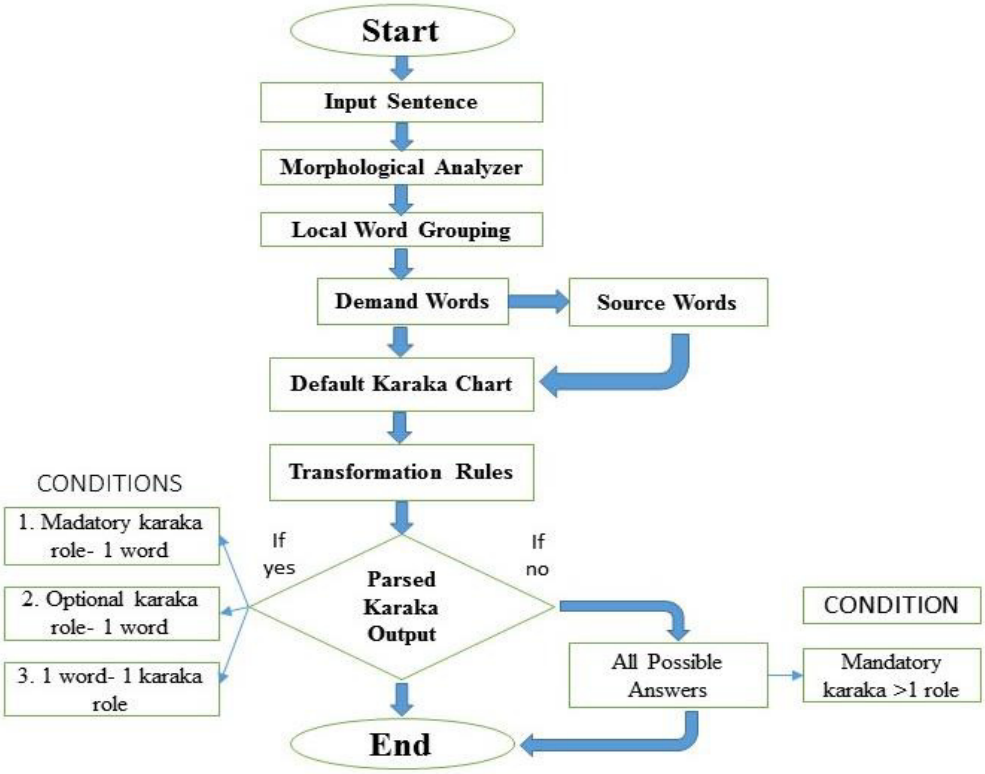


Fig. 2. the architecture of the kāraka parser

5. Conclusion

In the current study, we have focused on the errors with special reference to case in English-Urdu MT web-based platforms. It has been observed from the empirical data that sometimes the statistical MT platforms fail appropriately to have some of the case markers. Linguistically, we have proposed the kāraka-based PG dependency

analysis theoretical framework for the identification and resolution of kāraka-vibhakṯi/case markers errors. Computationally, we have further proposed an architecture of a parser based on the PG dependency for the automatic identification and parsing of semantic roles and postpositions.

References

1. A J Amita. An annotation scheme for English language using Paninian framework. *(IJISET)*: 2015; 2 (I): 616-619.
2. Bharati A & Sangal R. Parsing free word order languages in the Paninian framework. *ACL:* 1993; 105-111.
3. Bharati A, V Chaitanya, R Sangal & KV Ramakrishnamacharyulu. *Natural language processing: a Paninian perspective.* New Delhi: Prentice-Hall of India; 1995.
4. Bharati A, M Bhatia, V Chaitanya, R Sangal. Paninian grammar framework applied to English. Kanpur: *Department of Computer Science and Engineering, Indian Institute of Technology*; 1996.
5. Bharati A., D M Sharma, S Husain, L Bai, R Begam, R Sangal. Anncorra: treebanks for Indian languages, guidelines for annotating Hindi treebank. *IIIT Hyderabad*; 2009.
6. Bharati A., M Gupta, V Yadav, K Gali, D M Sharma. Simple parser for Indian languages in a dependency framework. *LAW*: *ACL*; 2009, 162-165.
7. Bhate S & Kak S. Pāṇini's grammar and computer science. *Annals of the Bhandarkar Oriental Research Institute*; 1991, 79-94.
8. Bhatia A, R Bhatt, B Narasimhan, M Palmer, O Rambow, D Sharma, M Tepper, F Xia. Empty categories in a Hindi treebank. *LREC*; 2010.
9. Bhatt R, B Narasimhan, M Palmer, O Rambow, D Sharma, F Xia. A multi-representational and multi-layered treebank for hindi/urdu. *LAW*: ACL; 2009, 186-189.
10. Bhatt R, O Rambow, F Xia. Linguistic Phenomena, Analyses, and Representations: Understanding Conversion between Treebanks. *IJCNLP: 2011,* 1234-1242.
11. Chaudhry H, H Sharma, D M Sharma. Divergences in English-Hindi parallel dependency treebanks. *DepLing; 2013*, 33.
12. Dey G & Maringanti HB. Paninian framework for odia language processing. 2014.
13. S C Kak. The Paninian approach to natural language processing. *International Journal of Approximate Reasoning*, 1987; I(1), 117-130.
14. Kiparsky P & Staal JF. Syntactic and semantic relations in Pāṇini.*Foundations of Language*. 1969; 83-117.
15. Muzaffar S, P Behera, GN Jha, L Hellan, D Beermann. The TypeCraft Natural Language Database: Annotating and Incorporating Urdu. INDJST; 2015, Vol 8(27), IPL0579. http://www.indjst.org/index.php/indjst/article/view/8 1728/63072
16. Muzaffar S & Behera P. Error analysis of the Urdu verb markers: a comparative study on Google and Bing machine translation platforms. *Aligarh Journal of Linguistics (ISSN- 2249-1511)*; 2015; 4 (1-2), 199-208.
17. Muzaffar S, Behera P, Jha GN. Issues and challenges in annotating Urdu action verbs on the Imagact4all platform. *LREC*; 2016.
18. Muzaffar S, P Behera, GN Jha. Classification and resolution of linguistic divergences in English-Urdu machine translation. *WILDRE: LREC;* 2016.
19. Rambow O. The simple truth about dependency and phrase structure representations: an opinion piece. *HLT: ACL*; 2010, 337-340.
20. Sharma DM, R Sangal, L Bai, R Begam, K V Ramakrishnamacharyulu,. AnnCorra: treeBanks for Indian languages. IIIT, Hyderabad, India; 2007.
21. Shukla P, D Shukl, A Kulkarni. Vibhakṯi divergence between Sanskrit and Hindi. In *Sanskrit Computational Linguistics*, Springer Berlin Heidelberg, 2010, 198-208.
22. RL Trask. *A dictionary of grammatical terms in linguistics*. Routledge; 2013.
23. Vaidya A, S Husain, P Mannem, D M Sharma. A Kāraka based annotation scheme for English. Springer Berlin Heidelberg; 2009, 41-52.
24. Behera P, Renu Singh, Girish Nath Jha. Evaluation of Anuvadaksh (EILMT) English-Odia Machine-assisted Translation Tool. *WILDRE: LREC;* 2016.