Issues and Challenges in Annotating Urdu Action Verbs on the IMAGACT4ALL Platform

Sharmin Muzaffar, Pitambar Behera & Girish Nath Jha

Aligarh Muslim University, Jawaharlal Nehru University

Centre for Linguistics

{sharmin.muzaffar, pitambarbehera2, girishjha}@gmail.com

Abstract

In South-Asian languages such as Hindi and Urdu, action verbs having compound constructions and serial verbs constructions pose serious problems for natural language processing and other linguistic tasks. Urdu is an Indo-Aryan language spoken by 51, 500, 000[1] speakers in India. Action verbs that occur spontaneously in day-to-day communication are highly ambiguous in nature semantically and as a consequence cause disambiguation issues that are relevant and applicable to Language Technologies (LT) like Machine Translation (MT) and Natural Language Processing (NLP). IMAGACT4ALL is an ontology-driven web-based platform developed by the University of Florence for storing action verbs and their inter-relations. This group is currently collaborating with Jawaharlal Nehru University (JNU) in India to connect Indian languages on this platform. Action verbs are frequently used in both written and spoken discourses and refer to various meanings because of their polysemic nature. The IMAGACT4ALL platform stores each 3d animation image, each one of them referring to a variety of possible ontological types, which in turn makes the annotation task for the annotator quite challenging with regard to selecting verb argument structure having a range of probability distribution. The authors, in this paper, discuss the issues and challenges such as complex predicates (compound and conjunct verbs), ambiguously animated video illustrations, semantic discrepancies, and the factors of verb-selection preferences that have produced significant problems in annotating Urdu verbs on the IMAGACT ontology.

Keywords: Action verbs; Compound verbs; Complex predicates; Conjunct verbs; Computational semantics; Semantic discrepancy; IMAGACT ontology.

1. Introduction

Action verbs are those which refer to activities and are the center of the predicate in any utterance or sentence. These verbs are the "most frequent structuring elements of the discourse" (Moneglia et al., 2012) in the communication process in any natural language of the world. Every language in the world categorizes actions in their own way which prevents the smooth functioning of NLP (Natural Language Processing), MT (Machine Translation) and other language technologies. Owing to the fact that they are 'polysemous' in nature (Moneglia et al., 2012); they often cause ambiguities that lead to various types of difficulties for the NLP technologies while processing.

Fig 1: IMAGACT Log-in Page

IMAGACT is a corpus-based ontology of action verbs which aims at setting up cross-linguistic ontology for disambiguation tasks in this crucial area of the lexicon (Moneglia et al., 2012) that deals with lexical semantics. It contains visual prototypes, but not definitions representing actions and thereby allows the exhibitions of typological variations across languages in transparent and informative manner. Figure 1 is a snapshot of the IMAGACT User log-in page.

1.1 The IMAGACT4ALL Platform

The IMAGACT data accounts for the semantic competence, separating the contexts from the metaphorical and the idiomatic expressions. Action verbs refer to a set of variety of possible ontological types which makes the task of the native speaker-annotator quite daunting with regard to selecting verb argument structure with a probability distribution. This paper discusses the issues and challenges- complex predicates (compound and conjunct verbs), ambiguously animated video illustrations, semantic discrepancies and the factors of verb-selection preferences- that have produced significant problems while annotating Urdu verbs on the IMAGACT ontology platform as native speakers of Urdu language. IMAGACT4ALL[2] is a 'competence-based extension' for the IMAGACT ontology. It provides wide representation of the actions that are most prominent in every-day life using prototypic 3d animations of brief forms.

IMAGACT is an online corpus of action verbs of various nature and the meta-languages initially used are English and Italian. So the prototypic animated video illustrations

[1] http://www.ethnologue.com/17/country/IN/languages/
[2] www.http://imagact.it/

are already explained in both the languages and the annotators of the other languages have to annotate their data based on the already-explained data in the meta-languages. Chart. 1 demonstrates a summary of verbs annotation in some major languages such as Chinese, English, and Italian. The number of verbs from Italian-English annotated is 515, translated is 473 out of the total 521. On the other hand, annotated and translated verbs from English-Italian are 546 and 497 in number respectively out of the total 550. So far as the annotation in the Chinese language is concerned, from Italian-Chinese 430 verbs have been annotated and 156 have been translated out of 521. In addition, the verbs from English-Chinese are 550 in number and the annotated and translated verbs are 30 and 22 respectively.[3]

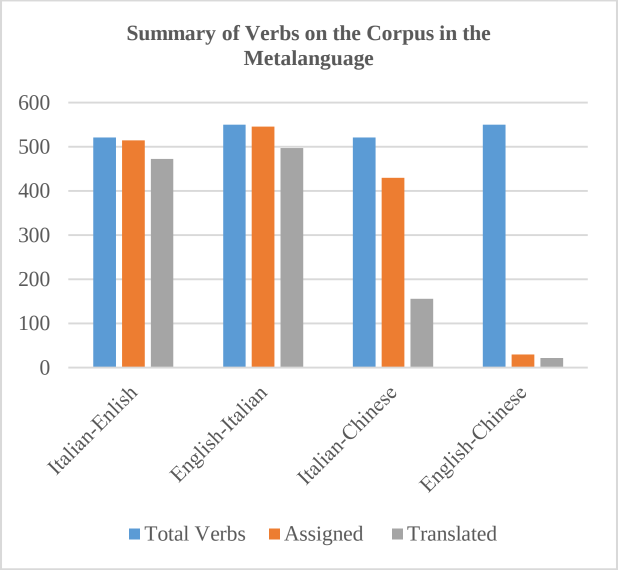


Chart 1: Summary of Verbs Annotation (till March, 2014)

It is currently being used for two main initiatives. The first one specifically concerns Indian languages: Bangla, Sanskrit, Hindi, Odia, Urdu, Manipuri and Magahi. A second initiative concerns languages like Polish, Danish, Tunisian and Tunisian-Arab and some others which are under the processing phase as the current extension beyond the existing Italian and English languages. The annotator needs to create these components, incorporating the applicable criteria that follow. The following chart represents the annotation of verbs in Indian languages. As far as Indian languages on the platform (Moneglia et al., 2014) are concerned, there are in totality 730 number of verbs that have already been annotated in seven Indian languages, viz., Bangla, Sanskrit, Hindi, Odia, Urdu, Magahi and Manipuri out of which 6 are scheduled languages excluding Magahi. In Bangla, Hindi and Odia, 110 number of verbs each has been annotated while the rest of the languages figures 100 verbs each.

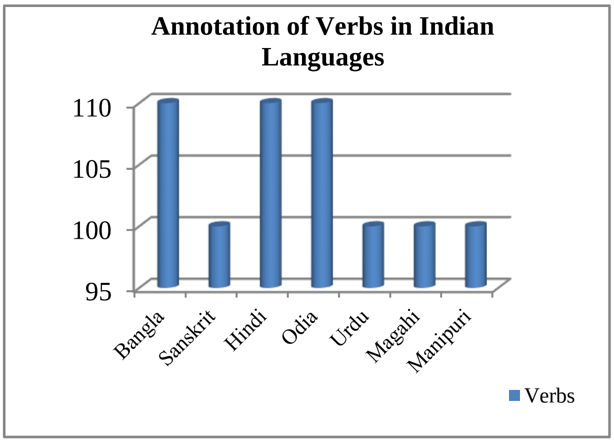


Chart 2: Annotated Verbs in Indian Languages on the IMAGACT

1.2 Urdu Language

Urdu, spoken approximately by 100 million people predominantly in Pakistan and India, is the member of the “Indo-Aryan language group within the Indo-European family of languages”.[4] It is the ‘national language’ of Pakistan (Rahman, 1996) along with being one of the two official languages. Further, it is also a Scheduled Language in the Union of India among other 22 languages. It is spoken all over the world owing to the fact that big South-Asian Diaspora is present in many parts of the world (Schmidt, 1999). It is also spoken in Bahrain, Afghanistan, Saudi Arabia, Bangladesh, Guyana, Botswana, Germany, India, Fiji, Malawi, Mauritius, Nepal, Oman, South Africa, Norway, Thailand, Qatar, the UAE, the UK, and Zambia.[5] Six Indian states have conferred Urdu an official status and one of the 22 Scheduled Languages in the Constitution of India.

Urdu has a few dialects, including Dakhni, Rekhta and Modern Vernacular Urdu (based on the Khariboli dialect of the western Uttar Pradesh region). It has a strong Perso-Arabic influence in its vocabulary, a cursive style, and context-sensitive Perso-Arabic script written from right to left and is closely associated to Hindi with which it shares commonalities with respect to morphology, syntax and a fair amount of vocabulary.

1.2.1. Forms of Urdu Verbs

As has been discussed in *Urdu: An Essential Grammar by* Ruth Laila Schmidt, Urdu verbs have been categorized into four main basic forms: root, imperfective and perfective participle and infinitive. They are as follows:

- Root

With the root /jA/ (go), /-nA/ suffix infinitive is attached to form an infinitive form /jAnA/ (to go) of the verb. Other roots vigorously used in the language are /kar/ (do), /de/ (give), sun (hear), /khA/ (eat) and the like.

- Imperfective Participle:

These participles are formed from the root by the addition

[3] http://www.imagact.it/imagact/cbeVerbsReport.seam

[4] www.ethnologue.com

[5] www.omniglot.com

of the present suffix /-tA/ /-te, -tI/, which is inflected like an adjective to agree with nouns or pronouns in gender and number.

/SunanA/, to hear- /sunatA/, hearing
/KaranA/, to do- /karatA/, doing
/KarAnA/, to cause to be done-/karAtA/, causing to be done
Forms of the imperfective participle:

	singular	plural
Masculine	/sunatA/	/sunate/
Feminine	/sunatI/	/sunatIṀ/

- Perfective Participles:

Perfective participles are formed from the root by the addition of the past suffix /-A, /-e, -I, -IṀ/, which is inflected like an adjective to agree with nouns or pronouns in gender and number. For instance, /SunanA/ 'to hear'-/sunA/ 'heard' /Pa.DhanA/ 'to read', /pa.Dha dēnA/ 'to read to someone' etc.

1.2.2. Types of Action Verbs in Urdu

- Simple and Compound Verbs:

Simple verbs are those which use only a single structure with no any compound structure. For instance, /sonA/ (to sleep) and /pa.DhanA/ (to read) are simple verbs while /sojAnA/ (to fall asleep) and /pa.Dha dēnA/ (to read to someone) are the examples of compound verbs (Agnihotri, 2007 and Schmidt, 1999). The simple verbs used in the IMAGACT platform are /TAnganA/ 'to put', /jo.DanA/ 'to fix', /DAlanA/ 'to put', /mo.DanA/ 'to roll up', /lu.DhakanA/ 'to roll', /belanA/ 'to roll', /lapeTanA/ 'to wrap', /ghUmanA/ 'to turn', /mu.DanA/ 'to turn', /palaTanA/ 'to turn', /laTakanA/ 'to hang', /karanA/ 'to do', /rakhanA/ 'to put', /baiThanA/ 'to sit', /bAMdhanA/ 'to bandage', /kATanA/ 'to cut', and /poMChanA/ 'to wipe' etc.

- Transitive and Intransitive Verbs

In transitive sentences the verb takes an object and the focus is on the action performed by the doer. On the contrary, intransitive verbs do not take objects and the emphasis is on the consequence of the action performed by the doer. English verbs can both be transitive and intransitive at the same point of time. For instance, the shopkeeper sells (transitive verb) fruits. The fruits are selling well in these days (intransitive verb). This phenomenon creates some problems for the annotator of the IMAGACT.

Example of a transitive verb:
nUra posTara mo.Da rahI hai
Noor-3.SG.FEM.NOM poster-ACC roll-PRES.IMPFV.PROG.AUX
"Noor is rolling up the poster."

In the above-instantiated example, the verb 'roll' takes up the object 'poster' as its argument and hence is a transitive one.

Example of an intransitive verb:
shAziA ghuma rahI hai
shazia- 3.SG.FEM.NOM spin- PRES.IMPFV.PROG.AUX
"shazia is spinning around."

In the above-mentioned example the verb 'to spin' does not take up any object and as a result can be called as intransitive verb.

On the IMAGACT platform the intransitive verbs employed are /lu.DhakanA/ 'to roll', /ghumanA/ 'to turn', and /palaTanA/ 'to turn' whereas the rest are transitive like /mo.DanA/ 'to roll up', /lapeTanA/ 'to wrap' and /laTakanA/ 'to hang'.

- Causative Verbs:

In Urdu causative verbs are formed by the addition of /-A/ /-la/ to the roots. For instance, from the root /sun/ and /khA/, causatives are formed like /sun-A/ and /khi-lA/. Double causatives are formed by the addition of /-vA/ and /-lvA/. For example, from the roots /kar/ and /khA/ causatives are formed like /kar-vA/ and /khil-vA/.

The causatives used on the IMAGACT platform are /saTAnA/ 'cause to set', /khisakAnA/ 'cause to move', /laTakAnA/ 'cause to hang', /phailAnA/ 'cause to lay', /miTAnA/ 'cause to rub', /ghumAnA/ 'cause to move', /hilAnA/ 'cause to stir', /cipakAnA/ 'cause to stick', /milAnA/ 'cause to mix', /sukhAnA/ 'cause to dry', /uchAlanA/ 'cause to toss', /lagAnA/ 'cause to lean', /jhukAnA/ 'cause to drop' , and /lu.DhakAnA/ 'cause to roll'.

2. Complex Predicates

The complex predicates are highly productive and different types can be stacked on top of one another (Butt, 2011), "so capturing their use computationally in a systematic, generalizable and efficient manner is a challenge" (Gunkel et al., 1998).

2.1 Conjunct Verbs

Conjunct verbs are those which comprise of noun or adjective and verb. In Hindi and Urdu Conjunct verb is formed by combining a noun or an adjective with a verb. They have the following structure (Begum et al., 2011)
Noun/adjective + verb = conjunct verb

The most frequent verbalizers in Hindi and Urdu are /karanA/ 'to do', /honA/ 'to be', /denA/ 'to give', /lenA/ 'to take', /AnA/ 'to come' (Begum, 2011).

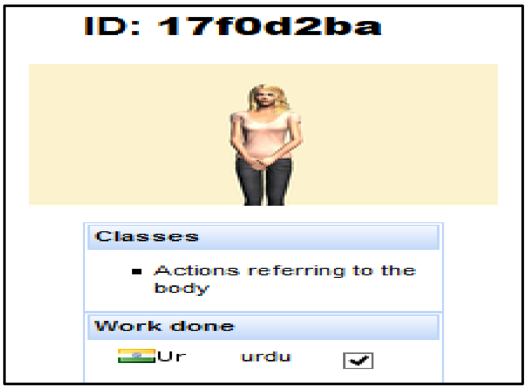
2.2Compound Verbs

Compound verbs are those which include a combination of verb with verb where the first verb in the occurrence is the polar verb and the second is the vector (Hook, 1974 and Butt, 2011). The vector explicates the semantic and grammatical aspects of the verb group and thereby the sentence (Abbi and Gopalakrishnan, 1991).
Verb (polar) + verb (vector) = compound verb

Some of the examples of compound verbs are /khA liyA/, /mAra DAlanA/, /de denA/, /toDa DiyA/ and so on.

3. Ambiguously Animated Video Illustrations

The animated videos as demonstrated in Figure 2 illustrated on the IMAGACT platform exhibit some ambiguity with reference to different types of actions. In other words, one action shows more than two types of representative verbs and hence causes disambiguation problems. As in the sentence id 17f0d2ba: /sir jhukAnA/ is the appropriate, /sir niche karnA/ is also possible, depicted by the picture.



ID: 17f0d2ba

Classes

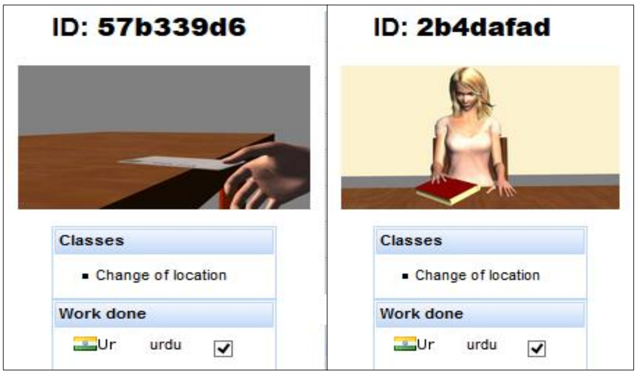
- Actions referring to the body

Work done

Ur urdu ☑

Fig 2: /Sir jhukAnA/ 'to drop head'

In both of the sentence id numbers 57b339d6, and 2b4dafad: the possible verb can be /ulaTanA/, which is not appropriate as it is used in the sense of putting some edibles from one object into some another. The correct verb is /palaTanA/ as it is illustrated in the afore-mentioned images that both the persons are turning the book and paper from the same location. In the sentence id no. b1be793, the picture depicts that she is dancing, but actually she is turning around.



ID: 57b339d6

Classes

- Change of location

Work done

Ur urdu ☑

ID: 2b4dafad

Classes

- Change of location

Work done

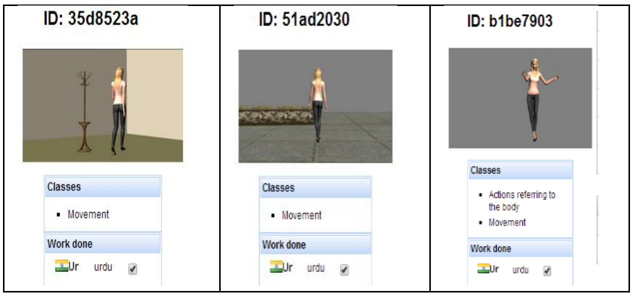
Ur urdu ☑

Fig 3: /palaTanA/ 'to turn over'

4. Semantic Discrepancy

In Figure 4, /ghumanA/ is used in all the ontologies despite the fact that in the id no.35d8523a it is used in the sense of moving while in the id no.51ad2030 it is employed as to refer to the sense of turning up. In the id no. b1be7903 it is used in the sense of spinning or rotating or turning. But in Urdu they are semantically treated as referring to the same action verb "moving" /ghumanA/.

On the contrary, in the English counterpart of the sentence ID: b65d7431ghumAnA refers to two verbs: 'to turn' and 'to rotate'. While in the rest of the sentence ids the verbs like rotate, swivel, turn and revolve are employed. It is because of the fact that each of the natural language assigns different action verbs for various ontologies and sometimes "one sole action verb in sentences having the same argument structure can refer to many different actions, so the verb does not explicitly specify the entity that it refers to" (Moneglia et al., 2012). This phenomenon of the language creates semantic discrepancy and is due to semantic factors. These action verbs are called general action verbs that can extend to actions belonging to different ontological types (Moneglia et al., 2012 and Mohanan, 1994). On one hand, there is no necessity for the existence of the object for ghumnA (to move) as it is an intransitive verb, whereas, on the other hand, it is likely for the action ghumAnA (cause to move) to take at least one object. In other words, in the causation of the action ghumAnA (cause to move) the action has to take the agent as the causal factor for the accomplishment of the action. Besides, another point which can be worthy of note is that it is also transitive verb along with causative.



ID: 35d8523a

Classes

- Movement

Work done

Ur urdu ☑

ID: 51ad2030

Classes

- Movement

Work done

Ur urdu ☑

ID: b1be7903

Classes

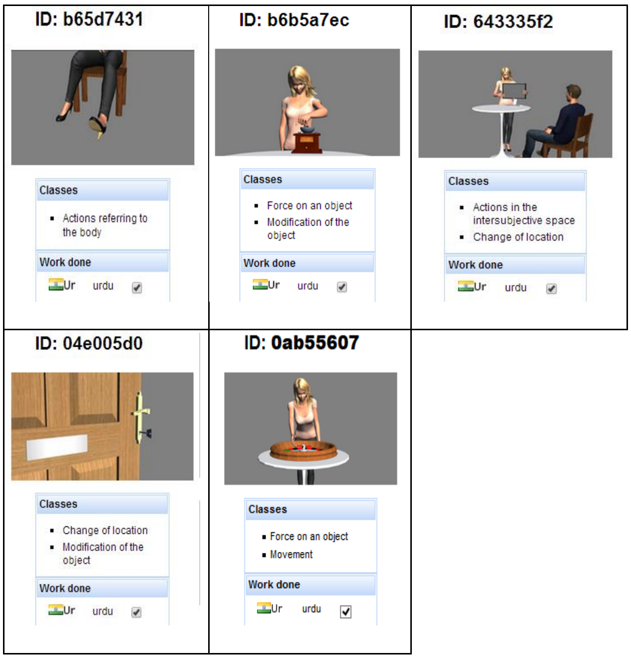
- Actions referring to the body
- Movement

Work done

Ur urdu ☑

Fig 4: /ghumanA/ 'to move'

Analogously, in the sentence id numbers such as cecbd89f and 2adb416f (see Figure. 5), /lu.DhakAnA/ is used and in the id f81899f2, /lu.DhakanA/ (roll) is used. But in the English counterpart for all the sentences, 'roll' is used to refer to all the ontological action verbs which reveals the above fact as averred by (Ishibashi, 2012 and Verma, 1999).



ID: b65d7431

Classes

- Actions referring to the body

Work done

Ur urdu ☑

ID: b6b5a7ec

Classes

- Force on an object
- Modification of the object

Work done

Ur urdu ☑

ID: 643335f2

Classes

- Actions in the intersubjective space
- Change of location

Work done

Ur urdu ☑

ID: 04e005d0

Classes

- Change of location
- Modification of the object

Work done

Ur urdu ☑

ID: 0ab55607

Classes

- Force on an object
- Movement

Work done

Ur urdu ☑

Fig 5: /ghumAnA/ 'to make something move'

5. Factors of Verb Selection Preferences

The verb 'to roll' in the sentence ids viz. cecbd89f and 2adb416f /lu.DhakanA/ is employed while the **id** f81899f2 contains lu.DhakanA is used in Urdu. The possible selection preferential candidate for the action of rolling could be /dhakelanA/ which is 'to push' for the said action.

The selection preference involves some factors: shape and weight of the object, force from the agent, direction of movement, causation and transitivity. In this case, the shape of the object is cylindrical and is light weight and the large amount of force or pressure exerted from the agent is indispensable. The participation of an agent and the patient is another criterion that the verb ought to take an object so as to fulfil the condition of transitivity, and finally the direction of movement is not a constraint which can be in any direction. The first two ids fulfill all the criteria mentioned and hence, are selected. The reason for /dhakelanA/ not being selected is that it involves the factors of direction of movement and the shape of the object. Generally, the shape of the object is a square-size and the movement has to be linear along a plain surface. The action /lu.DhakanA/ (see Figure. 6) also involves the agent and patient participation, and in the prototypic animated image the object is a cylinder and is rolling along a surface where the action verb /lu.DhakanA/ fits in.

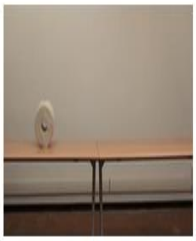
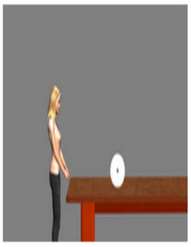
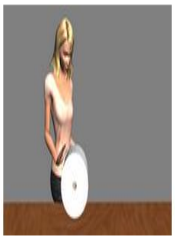
'Lu.DhakanA' to move	'Lu.DhakAnA' to cause something move	'Lu.DhakAnA' to cause something move
ID: f81899f2 	ID: cecbd89f 	ID: 2adb416f 

Fig 6: /lu.DhakanA/ vs /lu.DhakAnA/

6. Issues and Challenges

With the foregoing discussion, it can, however, be averred that working with IMAGACT4ALL encapsulates significant linguistic challenges with respect to verb selection in complex predicates (compound verbs and conjunct verbs), annotating ambiguous animated videos, factors for verb selection preferences and semantic discrepancy. The cultural difference regarding language poses serious problems for the annotator of a different cultural backdrop in order for appreciating the prototypic animated visuals and relate to his/her own socio-linguistic context. Single action verb in sentences having the same argument structure can denote to several types of actions. Therefore, the concerned action verb does not explicitly specify the entity that it signifies as put forth by (Ishibashi, 2012 and Verma, 1999). This above-stated phenomenon on the IMAGACT4ALL platform causes disambiguation problems. Moreover, the said platform has really attempted hard so as to avoid the intensity of the lexicographic work which upholds the idea of under-determinacy of semantic description. However, it has to be investigated up to which extent the animated video solves the issue of ambiguity; as the visuals capture the lexico-semantic aspect of the verbs. The association of prototypical visual scenes with the images provides a challenging question in restricting granularity to a minimal level as we cannot definitely specify the action suggested to by one action type as the example of another action verbs.

7. Conclusion

In this paper, the authors have dealt with the issues in annotating Urdu action verbs on the IMAGACT platform. If these issues and challenges are considered, the platform can further witness progress and not obstacles for incorporating more languages from across families; especially from families in Indian and the continent of Africa. One of the main NLP applications foreseen for IMAGACT4ALL platform is the word sense disambiguation. The objective of this paper correlates with the primary purpose of the said platform i.e. to incorporate more languages dealing with the pertinent issues. Thereby, it would be of a great value to the linguistic and NLP communities of different languages, especially less-resourced languages. The said ontological platform may further be of immense significance for Machine Translation and modelling of artificial intelligent systems. By way of incorporating less-resourced languages onto IMAGACT4ALL action ontology, international promotion and development of language resources and technologies may be achieved. Inter and trans-lingual research can be conducted between and among languages. Furthermore, languages having their genesis from a common parent language family can be compared considering verbs as the predominant category. We have emphasized the lexical-semantics and given less prominence to syntax. Furthermore, cross-familial comparison of action verbs can be initiated so as to map the ontological similarities and differences among the languages. Another important and interesting project could be comparing the cross-linking of the Indian Languages Corpora Initiative (ILCI) platform (Banerjee et al., 2013) which has seventeen Indian languages with the IMAGACT4ALL platform.

8. Acknowledgements

The IMAGACT project is funded by the PAR/FAS program of the Tuscan region in Italy.

9. Bibliographical References

Abbi, A. and Gopalakrishnan, D. (1991). Semantics of explicator compound verbs in south Asian languages. Language Sciences, 13(2), 161-180.

Agnihotri, R.K. (2007). Hindi: an essential grammar. London and New York: Routledge.

Alsina, A., Bresnan, J., & Sells, P. (1997). Complex predicates.

Banerjee, E., Kaushik, S., Nainwani, P., Bansal, A. and Jha, G. N. (2013). Linking and referencing multi-lingual corpora in Indian languages. Poland: 6th LTC Conference, 65-68.

Begum, R., Jindal, K., Jain, A., Husain, S. and Sharma, D. M. (2011). Identification of conjunct verbs in Hindi and its effect on parsing accuracy. Springer Journal.

Begum R., Jindal K., Jain A., Husain S., and Sharma D. M. (2011), Identification of conjunct verbs in hindi and its effect on parsing accuracy, Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, 29-40.

Butt, M. (2011). The light verb jungle: still hacking away. Complex predicates: cross-linguistic perspectives on event structure. 48-78.

Frontini, F., De Felice, I., Khan, F., Russo, I., Gagliardi, M. M. G., & Panunzi, A. (2012, December). Verb interpretation for basic action types: annotation, ontology induction and creation of prototypical scenes. In *24th International Conference on Computational Linguistics*, 69.

Haqqee, H. S. (2003). The Oxford English-Urdu Dictionary. USA: OUP.

Hook, P. E. (1991). The compound verb in Munda: An areal and typological overview. *Language Sciences*, *13*(2), 181-195.

Ishibashi, M. (2012). The expression of 'putting'and 'taking' events in Japanese. *Events of putting and taking: A crosslinguistic perspective*, *100*, 253.

Moneglia, M., Monachini, M., Calabrese, O., Panunzi, A., Frontini, F., Gagliardi, G., & Russo, I. (2012). The IMAGACT Cross-linguistic Ontology of Action. A new infrastructure for natural language disambiguation. In *LREC*, 2606-2613.

Moneglia, M. (2014). The variation of action verbs in multilingual spontaneous speech corpora. *Spoken Corpora and Linguistic Studies*, *61*, 152.

Moneglia, M., Brown, S.W. Kar, A., Kumar, A., Mello, H., Mishra, N., Jha, G.N., Ray, B. and Sharma, A. (2014). Mapping Indian Languages onto the IMAGACT ontology of action. *WILDRE*, Iceland.

M.K. Verma (ed.), Complex Predicates in South Asian Languages. New Delhi: Manohar Publishers and Distributors, 1999.

Mohanan, T. (1995). Wordhood and lexicality: Noun incorporation in Hindi. *Natural Language & Linguistic Theory*, *13*(1), 75-134.

Mohanan, T. (1994). Argument structure in Hindi. Stanford: CSLI Publication.

Muzaffar, S. and Behera, P. (2014). Error Analysis of the Urdu Verb Markers: A Comparative Study on Google and Bing Machine Translation Platforms, Aligarh Journal of Linguistics (ISSN- 2249-1511), 4 (1-2), pp 199-208.

Muzaffar, S., Behera, P. Jha G. N., Hellan L. & Beermann D. (2015). TypeCraft Natural Language Database: Annotating and Incorporating Urdu, Proceedings of the Third Reg-ICON, 2015. Indian Journal of Science and Technology, Vol 8(27), IPL0579. Accessed date 6.03.2016 http://www.indjst.org/index.php/indjst/article/view/81728/63072

Schmidt, R.L. (1999). Urdu: an essential grammar. London: Routledge.