“Taking Events” in Hindi. A Case Study from the Annotation of Indian Languages in IMAGACT

Massimo Moneglia, Alessandro Panunzi, Lorenzo Gregori
LABLITA – University of Florence
{massimo.moneglia, alessandro.panunzi, lorenzo.gregori}@unifi.it

Abstract

IMAGACT is a cross-linguistic ontology of action, in which action concepts are represented trough Prototypes (3D animations or brief films). The interface IMAGACT4ALL allows mother tongue informants to assign verbs of their language to each prototype and has been used to implement languages belonging to different families. The Ontology specifies the range of different actions which may fall in the extension of each action verb and the set of verbs which can identify each entry, ensuring an adequate translation to action verbs, which show high ambiguity and cross-linguistic semantic variability. A large initiative for the implementation of various Indian languages (Hindi, Urdu, Sanskrit, Bengali, Odia, Assamese, Magahi, Manipuri, Tamil) was undertaken. The paper sketches the status of the work, whose main achievement is the full implementation of Hindi/Urdu and focus on “taking events”, that are very relevant in ordinary communication, but feature strong differences in lexical encoding cross-languages. Hindi requires 7 different verbs to cover the actions extended by the general verb take. The main translator लेना(lenA) is also a general verb, but its application has specific semantic boundaries. The paper specifies how features are induced from prototypes, exploiting IMAGACT for the semantic interpretation of Hindi verbs.

Keywords: Action Ontology, Verb Semantics, Comparative Semantics

1. Indian Languages in IMAGACT

IMAGACT is a cross-linguistic ontology of action, in which the entries are prototypic 3D animations (or brief films), each one representing a distinct action concept. Concepts in IMAGACT are connected to a wide set of action verbs with strong impact in the language use: the selected verbs are the ones with highest frequency in speech corpora (Moneglia, 2014; Moneglia and Panunzi, 2007). The ontology of Action (1,010 concepts in the first release) has been induced through a controlled methodology (Moneglia et al., 2012) from English and Italian spoken corpora (Moneglia et al., 2014), grounding relevant concepts on the actual actions referred therein.

The outcome of the induction process leads to specify the set of Italian and English verbs which can be used to refer to each action prototype.

The use of images for action concepts identification allows to extend the «Verb(s)-Action prototype» correlation to any language through competence-based judgments. The web interface IMAGACT4ALL has been designed to allow mother tongue informants to assign verbs of their native language to each entry. Once mapped onto the ontology, each language can be compared to the others. More specifically, the appropriate verb(s) for each action entry (in every implemented language) is specified and the range of action concepts extended by each verb can be compared to the other within and across languages. IMAGACT is therefore a mean to make clear how languages convey a specific semantic categorization of action and also a mean to assist the translation process, specifying what are the verbs required by a given language to identify each particular action type.

IMAGACT have been extended to Chinese and Spanish (Brown et al., 2014) and to a set of languages of different families: Slavonic Languages (Polish and Serbian), Romance languages (Portuguese), German Languages (German and Danish), Arabic and Japanese. Moreover a specific campaign for implementing Indian languages has been undertaken (Moneglia et al., 2014). So far nine languages belonging to three language families has been considered: Sino-Tibetan (Manipuri), Dravidian (Tamil) and Indo-Aryan (Sanskrit, Hindi, Urdu, Odia, Bengali, Assamese, Magahi).

Table 1 specifies the number of processed entries in the Ontology and the number of Action verbs recorded for each Indian language under processing.

Language	Processed scenes	Verbs	Average Scenes per Verb
Assamese	150	103	1.46
Bangla	260	246	1.48
Hindi	1,006	512	2.39
Magahi	100	68	1.59
Manipuri	100	64	1.56
Oryia	110	178	1.28
Sanskrit	212	292	1.83
Tamil	100	95	1.19

Table 1: Number of processed scenes, inserted verbs and the average number of scenes per verb.

Issues and challenges regarding Urdu action verbs have been discussed by Muzaffar et al. (2016). Behera et al. (2016) focused on the possible benefit of this data base for translation. The full implementation of Hindi / Urdu in IMAGACT is a crucial milestone and creates now the possibility of large scale comparison with the other languages and in particular with English. Here we will specifically consider the value of this resource for making objective the peculiar semantic feature which characterizes Hindi verbal lexicon referring to action.

The semantic side is crucial for language disambiguation and translation. There is no one to one correspondence between action concepts and verbs. The number of verbs which can identify one Action may vary from language to language and one verb can in turn identify many different

actions. We call “General” those verbs which share this property.

Verb	Num. Scenes	Verb	Num. Scenes
लगाना (lagAnA)	38	मिलाना (milAnA)	11
रखना (rakhanA)	33	काटना (kATanA)	10
खोलना (kholanA)	30	गिरना (giranA)	10
निकालना (nikAlanA)	24	उतरना (utaranA)	10
उठाना (uThAnA)	21	लाना (lAnA)	9
डालना (DAlanA)	19	पलटना (palaTanA)	9
खींचना (khIMcanA)	18	भरना (bharanA)	9
हटाना (haTAnA)	15	फैलाना (phailAnA)	9
बंद करना (baMda karanA)	14	देना (denA)	9
मारना (mAranA)	13	ले जाना (le jAnA)	9
तोड़ना (to.DanA)	13	छोड़ना (cho.DanA)	8
दबाना (dabAnA)	12	हिलाना (hilAnA)	8
फेंकना (pheMkanA)	12	घुमाना (ghumAnA)	8
बांधना (bAMdhanA)	12	लेना (lenA)	7
गिराना (girAnA)	11	लपेटना (lapeTanA)	7
बंद करना (baMda karanA)	11	खिसकाना (khisakAnA)	7
जोड़ना (jo.DanA)	11	पकड़ना (paka.DanA)	7
चलाना (calAnA)	11		

Table 2: The first 35 general action verbs in Hindi.

Hindi, like English and Italian, characterizes for the presence of many verbs which can be applied to many different Action Concepts (Moneglia et al., 2014). Table 2 specifically presents the Hindi action verbs which can be interpreted according to the larger variety of different prototypes.

This paper is dedicated to the induction of semantic properties of a highly ambiguous language concept, the ones related to «taking events». We will show how a process of semantic feature extraction can be performed starting from IMAGACT. Prototypes and how the procedure should be driven by the annotations which IMAGACT makes available.

2. Taking events in English and Hindi

2.1 The variation of Take across Action Types

One action verb like *to take* is understood by competent speakers as one single action. However, as for many high frequency verbs, it does not refer to a unique action concept, but to many different concepts in the actual language usage. The Figure 1, derived from IMAGACT, shows this phenomenon. The set of prototypes identify how *take* vary its possible reference across different action concepts.

The typological distinction among actions in the extension of one general verb is supported by the fact that different verbs with different meaning are able to identify the same action. Looking to Figure 1, almost each action prototype feature one or more local equivalence with other action verbs, like *to extract*, *to receive*, *to remove*, *to bring*, *to lead*, *to grasp* and so on. This equivalence marks the difference among the represented actions and constitute an explicit differential of each concept prototype with respect to the others.

Once the range of relevant variations and their differential is identified, concepts can be modelled and generalizations obtained. For instance, the set of actions extended by *to take* fall in a restrict set of models roughly

Figure 1: The variation of *to take* across Action Types

identified by a higher level local equivalence (*to remove*, *to receive*, *to bring* and *to grasp*). In conclusion, there are many types of *taking* events which fall under the extension of *to take* and they can be gathered into classes according to high level local equivalence variations, designing the language specific categorization of *taking* events into English.

We do not know exactly what are the boundaries which limit the possible variation of a verbal entry referring to "taking events", however, putting this question at cross-linguistic level, we can see that each language parse the continuum in its own way (Kopecka and Narasimhan, 2012) and starting form IMAGACT data we can make objective what are the differentials among languages.

2.2 Taking events in Hindi

There is not an Hindi verb which covers the full range of applications of *to take*: 7 different verbs are recorded in IMAGACT to satisfy the variation of the English verb, respectively लेना (lenA), पकड़ना (paka.DanA), उठाना (uThAnA), हटाना (haTAnA), निकालना (nikAlanA), लाना (lAnA), ले जाना (le jAnA).

The main translator, लेना (lenA), applies to those taking activities in which the goal is that the "object comes in possession of the agent". This feature can be induced from the small selection of prototypes extended by लेना (lenA) in IMAGACT, compared to the large variation of *to take*. Figure 2 shows a selection of prototypes where both the predicates can be applied face to those that are extended by *take* only (on the right).

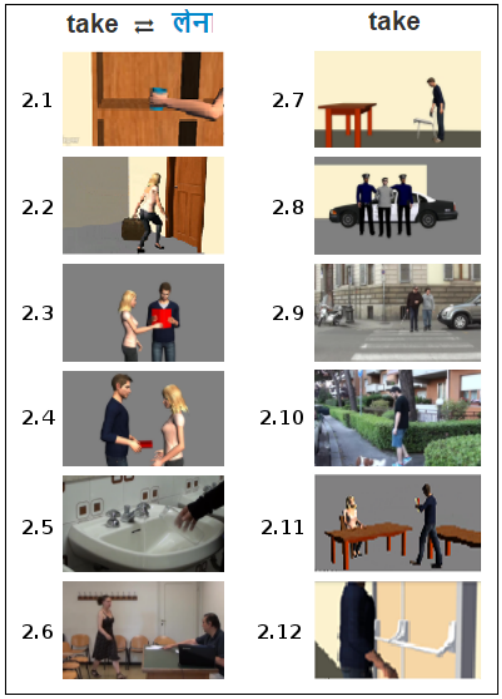

Figure 2: Comparison *take* vs लेना (lenA)

The resulting state "object in possession of the Agent" occurs in all prototypes in which *take* / लेना (lenA) are equivalent, i.e. when "getting object from its location" (2.1), when "getting and bringing the object" (2.2), when "taking is privative of somebody" (2.3), when "taking is also receiving from somebody" (2.4-2.6) or "from a source" (2.5). Under this semantic assumption, it is straightforward the conclusion that the meaning of लेना (lenA) is not appropriate to identify events in which *to take* is equivalent to *to bring* (2.7), *to carry* (2.7), *to lead* (2.8; 2.9; 2.10) and *to give* (2.11), in which the object necessarily have other destinations than the agent.

In parallel, we also find a reason why *grasping* events (2.12) are not extended by लेना (lenA), since the object in these event is "handled", but does not come in the possession of the agent. IMAGACT shows that the verb पकड़ना (paka.DanA) is appropriate in this case (see below).

Those taking events in Figure 1 which the object is extracted from a container or raised from a lower position are respectively captured by the specific predicates उठाना (uThAnA) (to *pick-up / to rise*) and निकालना (nikAlanA) (*to remove / to extract*) (see Figures 3 and 4). However, those events may be also extended by लेना (lenA), which behave as local equivalent of this verb. Indeed, for getting in the possession of an object, we frequently rise or extract it from its collocation[1].

Figure 3: Comparison *take* / उठाना (uThAnA)

IMAGACT makes clear the local nature of the relation of उठाना (uThAnA) and निकालना (nikAlanA) with लेना

[1] It remains unclear from IMAGACT data what are the limits of this equivalence. In some taking prototypes, the object indeed is raised, but उठाना (uThAnA) is not marked in the annotation. The same is when getting an object from a container निकालना (nikAlanA). According to a close evaluation of IMAGACT data for this equivalence relation, the application of लेना (lenA) is possible all the time the object come in the possession of the agent, although the event is categorized by preference with specific predicates. Thanks to Atul Ojha for providing data on this issue.

(lenA). For instance, उठाना (uThAnA) extends to a lot of events where no act of taking is performed. Figures 3 and 4 show the essential of the comparison between the two predicates and taking events. On the right side is displayed the large set of prototypes in which the object is raised or extracted, but no taking event occurs.

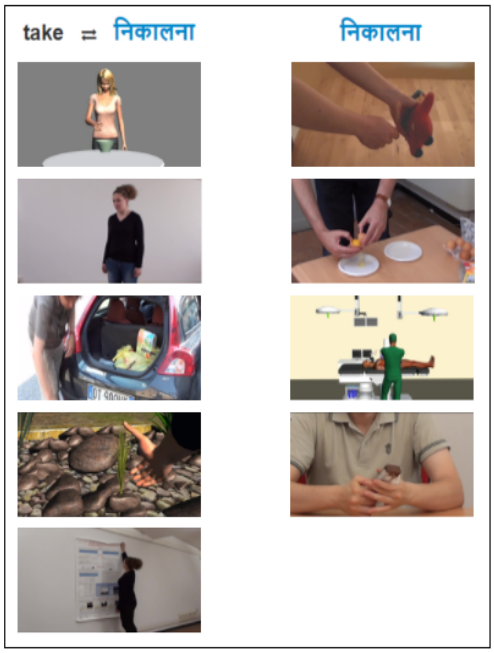

Figure 4: Comparison *take* / निकालना (nikAlanA)

Ragarding taking events where the English verb is equivalent to *remove* and/or *extract*, we can notice that the focus of the taking activity is not the «coming in control of the object», but rather that the object loses its original collocation. This may be the reason why in IMAGACT those prototypes are not marked as the extension of लेना (lenA), which is however marginally acceptable. The appropriate Hindi verbs are respectively निकालना (nikAlanA) (*to extract*) and हटाना (haTAnA) (*take away*).

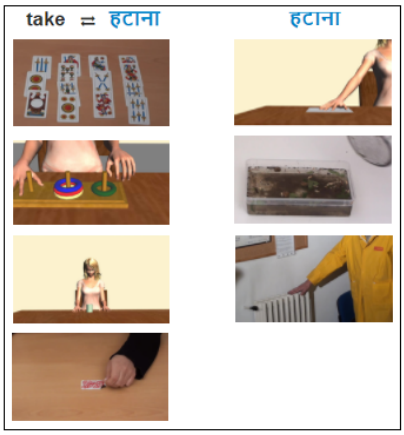

Figure 5: Comparison *take* / हटाना (haTAnA)

Looking at glance to IMAGACT data, the induction of differential semantic features from these prototypes is immediate. Hindi closely distinguish “extractions” from “displacements events”. Indeed, as the comparison in Figure 6 shows, the intersection between हटाना (haTAnA) and निकालना (nikAlanA) in IMAGACT is limited to the events in which displacement is reached through extraction (i.e *extract/remove* a substance from a liquid).

Figure 6: Displacements हटाना (haTAnA) *vs* Extractions निकालना (nikAlanA)

IMAGACT does not gives alternatives to these verbs. It seems that Hindi prefer specific verbs to the general verb लेना (lenA), when removal events take place.
Grasping events are identified by the Hindi verb पकड़ना (paka.DanA), which covers the fields of application where *to take* is equivalent to *to grasp* and *to grab* (Figure 7).

Figure 7: The variation of पकड़ना (paka.DanA)

The range of extensions of the Hindi verb, however, is larger and it over-extends with respect to the range of applications of *to take*, covering also «catching events», that cannot be identified by *take*. The extension of पकड़ना (paka.DanA) to the fields of application of *to catch* is not surprising. For instance the general verbs

coger in Spanish and *prendere* in Italian, can also refer to catching events in local equivalence with other specific verbs (respectively *agarrar* and *acchiappare*).
Contrary to English (and Arabic), bringing events cannot be in the extension of any general Hindi verb referring to the *reaching*, *grasping*, *taking* sequence. Looking to the English variation, in order to predicate of bringing events, the set of equivalent verbs available in the place of *to take*, specifies at least four categories: 1) *bringing / moving* (partially overlapping displacement types); 2) *bringing / giving*; 3) *bringing / carrying*; 4) *bringing / leading*. IMAGACT specifies that Hindi applies two verbs to the events in this variation, respectively लाना (lAnA) and ले जाना (le jAnA).

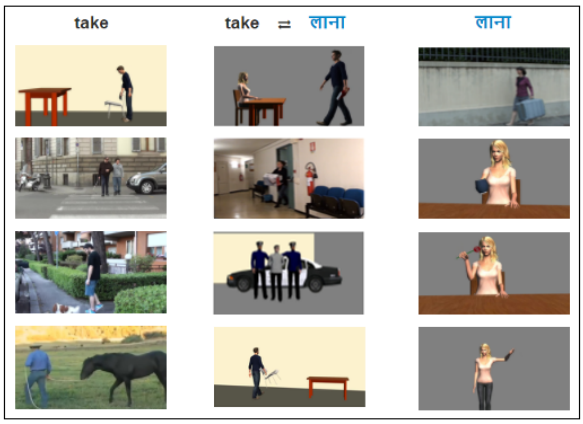

Figure 8: Comparison of *take* / लाना (lAnA)

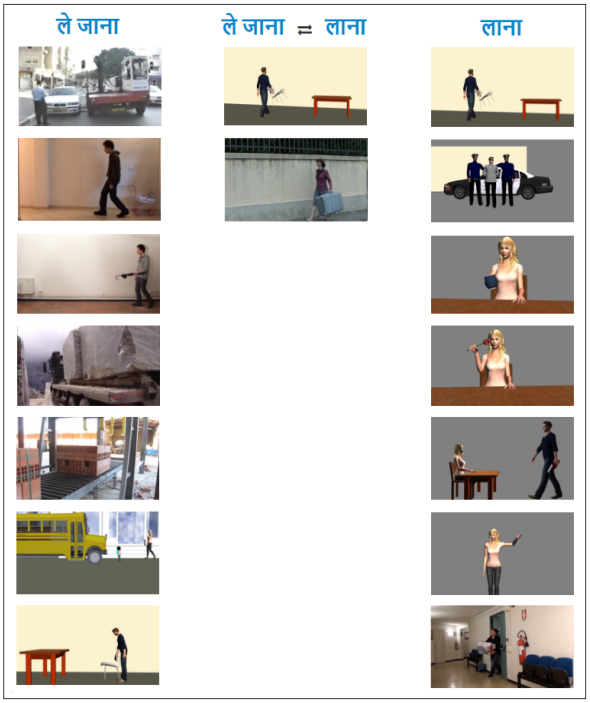

Figure 9: Comparison of ले जाना (le jAnA) *vs* लाना (lAnA)

Figure 8 shows that लाना (lAnA) is quite general, since it can be applied to three categories of bringing events: "bring/give", "bring/move", "bring/carry" (in the centered column), and only shows restrictions on some "bring/lead" events. As Figure 8 also shows, लाना (lAnA) over-extends taking events, since it refers in general to the act of bringing, both when carrying an object in space and when moving an object to a position (right column). ले जाना (le jAnA) partially overlaps लाना (lAnA). As the comparison in Figure 9 shows, both verbs can be applied when movement in space by the subject is accompanied with holding one object (centred column), but ले जाना (le jAnA) appear specifically appropriate to transportation (on the left column) and, contrary to लाना (lAnA), it does not extend to events in which the object is not carried but just moved (right column).
Among the set of action types extended by *take* in Figure 1 IMAGACT does not provide clear results for the identification of leading / guiding events in Hindi, marking with different Hindi verbs similar prototypes, whose differentials are not evident to the user for feature extraction.

Figure 10: Leading events in Hindi

3. Conclusions

The translation of *take* into Hindi and on the other way around the translation in English of the various Hindi verbs that are needed to cover the set of events falling into the *reaching*, *grasping taking* sequence, requires a clear pragmatic knowledge. Verbs are not in translation relation among them, but find their correspondence in specific types of activities. Looking to the set of Action types which falls within the extension of each concerned verb, according to the variation provided by IMAGACT, we figured out that cross-linguistic correspondences follow from semantic regularities. In so doing we have shown that the IMAGACT infrastructure can be used as a core source of information for the semantic modelling of Indian Languages, which, like Hindi, can be compared with other languages on the basis of explicit semantic knowledge.

4. Acknowledgments

The annotation of Hindi has been achieved by prof. Girish Nath Jha who also processed Urdu in collaboration with Sharmin Muzaffar. We also thanks Atul Ojha and the following students for helping in the processing of IMAGACT entries: Himani, Shivek, Shagun, Geeta, Zoya for Hindi; Debmalya for Bangala; Abhijit for Sanskrit; Rajamatangi and Selva for Tamil; Prachi for Marathi; Diksha, Bimrisha and Alvina for Assamese.

5. Bibliographical References

Kopecka, A. and Narasimhan, B. (2012). Events of Putting and Taking, A Cross-linguistic Perspective. Amsterdam: Benjamins.

Moneglia, M., Brown, S. W., Frontini, F., Gagliardi, G., Khan, F., Monachini, M. and Panunzi, A. (2014). The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the Representation of Lexical Encoding of Action. In Nicoletta Calzolari et al., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 3425-3432. European Language Resources Association (ELRA).

Moneglia M. (2014). Natural Language Ontology of Action: A Gap with Huge Consequences for Natural Language Understanding and Machine Translation. In Vetulani, Z. and Mariani, J., editors, Human Language Technology Challenges for Computer Science and Linguistics, Lecture Notes in Computer Science 2014, 5th Language and Technology Conference (LTC 2011), pages 370-395. Springer.

Moneglia, M., Gagliardi, G., Panunzi, A., Frontini, F., Russo, I., and Monachini, M. (2012). IMAGACT: Deriving an action ontology from spoken corpora. Paper presented at the Eight Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-8). Pisa, October 3-5, 2012.

Moneglia, M., Brown, S. W., Kar, A., Kumar, A., Ojha, A. K., Mello, H., Niharika, Nath Jha, G., Ray, B. and Sharma, A. (2014). Mapping Indian Languages onto the IMAGACT Visual Ontology of Action. In 2nd Workshop on Indian Language Data: Resources and Evaluation (WILDRE-2), pages 51-55.

Muzaffar, S., Behera, P. and Nath Jha, G. (2016). Issues and Challenges in Annotating Urdu Action Verbs on the IMAGACT4ALL Platform. In Nicoletta Calzolari et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1446-1451. European Language Resources Association (ELRA).

Brown, S. W., Gagliardi, G. and Moneglia, M. (2014). IMAGACT4ALL: Mapping Spanish Varieties onto a Corpus-Based Ontology of Action. *CHIMERA* 1:91-135.

Behera, P. Muzaffar, S., Ojha, A. K. and Nath Jha, G. (2016). The IMAGACT4ALL Ontology of Animated Images: Implications for Theoretical and Machine Translation of Action Verbs from English-Indian Languages. In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2016), pages 64-73.

6. Language Resource References

IMAGACT. http://www.imagact.it