

# One event, many representations. Mapping action concepts through visual features.

Alessandro Panunzi, Andrea Amelio Ravelli, Lorenzo Gregori

Università degli Studi di Firenze, Florence, Italy  
{alessandro.panunzi, lorenzo.gregori, andreaamelio.ravelli}@unifi.it

## Abstract

This paper faces the problem of unifying the representation of actions and events in different semantic resources. The proposed solution exploits the IMAGACT visual component (video scenes that represent physical actions) as the linkage point among resources. By using visual objects, we connected resources responding to different scopes and theoretical frameworks, in which a concept-to-concept mapping appeared difficult to obtain. We provide a brief description of two experiments that exploit IMAGACT videos as a linkage point: an automatic linking with BabelNet, a multilingual semantic network, and a manual linking with Praxicon, a conceptual knowledge base of action. The aim of this work is to integrate data from resources with different level of granularity in order to describe the action semantics from a linguistic, visual and motor point of view.

**Keywords:** ontology linking, semantics, action verbs, knowledge representation

## 1. Introduction

Action verb interpretation is a basic issue for human-machine interaction systems that aim to process natural language instructions. The difficulty behind automatic action verb understanding comes out from the evidence that no one-to-one correspondence can be established between action predicates (lexical items in each natural language) and action concepts (mental representations of experienced events). The same action can be predicated by multiple verbs and, conversely, one verb can extend to multiple and different actions. Most of these verbs belong to the class of general verbs, which are characterized by a high ambiguity and high frequency in the use (Moneglia, 2010). In these circumstances, senses are often vague and overlapping, their discrimination is not clear, and this is a critical issue for their semantic representation. In fact, if we look at WordNet (Fellbaum, 1998), the proposed classification of general action verbs highlights two main issues: on one hand, a synset often encodes a variety of events that are cognitively conceived as different action concepts; on the other hand, it's frequent that a specific event is not clearly described by a unique verb sense, but it seems to be spread on more senses (belonging to different synsets), each one representing a possible conceptualization of the same event. The following examples explain these classification problems.

Example 1 shows a synset that represents a very general sense of *putting objects in a location* that can refer to actions of pouring, inserting or laying a body part. In the Example 2 there is an action of *beating up someone* (Figure 1), that can be correctly encoded by two senses of the verb *to beat*.

### Example 1

**bn:00090224v<sup>1</sup>:** place, lay, put, set, pose. *Put into a certain place or abstract location.*

<sup>1</sup>The IDs reported in Example 1 and 2 are taken from BabelNet taxonomy, that derives directly from WordNet. (see 2.2. for more details.

- John puts the wine in the glass;
- John puts the letter in the envelop;
- John puts his hand on Mary's shoulder.

### Example 2

- **bn:00083248v:** beat up, beat, work over. *Give a beating to; subject to a beating, either as a punishment or as an act of aggression;*
- **bn:00083249v:** beat. *Hit repeatedly.*



Figure 1: The action *John beats/beats up/batters Paul*.

The action representation is even more difficult in a multilingual perspective, given that different languages operate different segmentations of the action domain; for example, classification methods built upon English language do not necessarily hold in other languages, and especially for typologically-different language families (Majid et al., 2007). It has been observed that often it is not possible to find an exact match between lexicalized action concepts in different languages, even with a fine-grained sense distinction (Moneglia and Panunzi, 2007). Moreover, one language could totally lack a lexical representation for a specific concept, whenever there is a lexical gap (Gregori and Panunzi, 2017).

These considerations highlight a big issue in the creation of linguistic ontologies, where there is a need to divide the word meaning in a set of senses that are discrete, well-defined and related together through a predefined set of semantic relations. A word sense discrimination task is tricky and the existence of an universal set of word senses is questioned both theoretically (Wittgenstein, 1953; Pustejovsky, 1991; Croft and Cruse, 2004) and computationally (Kilgarriff, 1997; Resnik and Yarowsky, 1999; Cimiano et al., 2013).

The difficulty in finding a shared representation of concepts reflects in the existence of a wide variety of ontologies and lexical resources that are often bounded to a specific theoretical model and have different levels of granularity in their concept definition. In this context, finding matches between concepts encoded in different resources is a hard task.

We describe here a *visual mapping* methodology, that has been applied to connect together action concepts from different resources. Instead of a classic concept-to-concept mapping, visual mapping performs a concept-to-video linking. In fact, a video depicting an event is not subject to any linguistic constraint, and therefore the associated semantic information can be described in various manners. Starting from this observation we used videos to link concepts of different resources, that express independent event conceptualization according to their own theoretical framework.

By exploiting the videos featured in the IMAGACT ontology of action, we applied the visual mapping to connect BabelNet, a general multilingual semantic network, and Praxicon, a specific conceptual knowledge base of action.

## 2. Resources

### 2.1. IMAGACT

IMAGACT Visual Ontology of Action<sup>2</sup> (Moneglia et al., 2014) is a multimodal and multilingual resource that offers a novel integration of visual and linguistic information as complementary elements. The resource contains 1010 distinct action concepts as a result of an information bootstrapping from Italian and English spoken corpora. Metaphorical and phraseological usages have been excluded from the annotation process, in order to collect only the occurrences referring to physical actions.

Verbs in IMAGACT are divided into *action types*, according to their semantic variation. An action type gathers a group of actions, that are perceived as unitary from a cognitive point of view. Each type is linked to one or more video scenes (either 3D animations or filmed video clips) of performed actions, that act as prototypes for it. The verbs of each language referring to the same actions are linked to the same scenes, resulting in an interlinguistic and multimodal semantic network.

The ontology is in continuous development and, at present, contains 9 languages and 13 more that are under development, with an average of 730 action verbs per language.

This resource gives a broad picture of the variety of actions and activities that are prominent in everyday life and

<sup>2</sup><http://www.imagact.it/>

specifies the lexicon used to express each one in ordinary communication, in all the included languages.

### 2.2. BabelNet

BabelNet<sup>3</sup> (Navigli and Ponzetto, 2012) is a multilingual semantic network developed through the automatic mapping of the WordNet thesaurus and the Wikipedia encyclopedia. At present, BabelNet 3.7 contains 284 languages and it is the widest multilingual resources available for semantic disambiguation. Concepts and named entities are represented by BabelSynsets (BSs), unitary concepts identified by several kinds of information (semantic features, glosses, usage examples, images, etc.) and related to lemmas (in any language) which have a sense matching with that concepts. BSs are not isolated, but connected together into a huge network by means of the semantic relations inherited from WordNet.

### 2.3. Praxicon

Praxicon<sup>4</sup> is an ontology for the representation of action concepts, based on the Minimalistic Grammar of Action (Pastra and Aloimonos, 2012). In Praxicon, an action is expressed through motor concepts, specified in terms of 3 basic components: GOAL, TOOL and OBJECT. A wide part of this ontology is also linked with WordNet synsets and ImageNet images (Deng et al., 2009).

Praxicon makes a distinction between *Actions*, *Movements*, and *Events*<sup>5</sup>. Actions are sets of structured motor execution, intentionally performed by an agent with a tool to achieve a goal. The goal is a necessary component, so any non-voluntary motor activation is addressed as a Movement, but not as an Action. Finally, actions that are too complex to be described as a set of motor concepts, are considered Events and are out of the scope of the Praxicon resource.

## 3. Visual mapping at work

Herein we show how the visual mapping technique has been applied to link IMAGACT with BabelNet and Praxicon.

An example of the linking between IMAGACT, Praxicon and BabelNet can be observed in Figure 2, that shows a *beating* event with the parallel representation in the 3 resources.

### 3.1. IMAGACT and BabelNet



BabelNet concepts (the BSs) are interlinguistic: they gather all the word senses in different languages that are semantically equivalent (or almost equivalent). Conversely, IMAGACT action types encode small semantic differences, so they are more granular and language-dependent. Given these differences, an exact match between concepts is very rare; it's also hard to establish less strict semantic relations (e.g. *narrow-to-broad*), because the BSs boundaries are often fuzzy and the gloss is not always able to make a clear discrimination between them.

<sup>3</sup><http://www.babelnet.org/>

<sup>4</sup><https://github.com/CSRI/PraxiconDB>

<sup>5</sup>These categories have their own definition in the Praxicon framework. We use capital letters when referring to this specific meaning

Figure 2: An example of the resulting linking between BabelNet, IMAGACT and Praxicon.

BABELNET		BabelSynset: bn:00083248v Gloss (WordNet): Give a beating to; subject to a beating, either as a punishment or as an act of aggression	IT: bastonare, battere, malmenare, percuotere, <b>picchiare</b> EN: <b>beat up, beat</b> , work over SP: apalear, dar una paliza, derrotar, <b>golpear, pegar</b> ZH: 殴打, 打 [...] [...]						
IMAGACT	IT: pestare, <b>picchiare</b> , menare, battere, colpire EN: strike, hit, <b>beat</b> , batter, <b>beat up</b> SP: agredir, <b>golpear, pegar</b> ZH: 打, 揍, 拳打脚踢 [...] [...]								
	scene ID: 4faba0b4 								
PRAXICON	<table border="1" style="width: 100%; text-align: center;"> <tr> <td>[punch]<sub>GOAL</sub> [with fist]<sub>TOOL</sub> [the chest]<sub>OBJECT</sub></td> <td>[punch]<sub>GOAL</sub> [with fist]<sub>TOOL</sub> [the stomach]<sub>OBJECT</sub></td> <td>[hit]<sub>GOAL</sub> [with elbow]<sub>TOOL</sub> [the back]<sub>OBJECT</sub></td> <td>[hit]<sub>GOAL</sub> [with knee]<sub>TOOL</sub> [the back]<sub>OBJECT</sub></td> <td>UNINTENTIONAL {NO MATCH}</td> </tr> </table>				[punch] <sub>GOAL</sub> [with fist] <sub>TOOL</sub> [the chest] <sub>OBJECT</sub>	[punch] <sub>GOAL</sub> [with fist] <sub>TOOL</sub> [the stomach] <sub>OBJECT</sub>	[hit] <sub>GOAL</sub> [with elbow] <sub>TOOL</sub> [the back] <sub>OBJECT</sub>	[hit] <sub>GOAL</sub> [with knee] <sub>TOOL</sub> [the back] <sub>OBJECT</sub>	UNINTENTIONAL {NO MATCH}
[punch] <sub>GOAL</sub> [with fist] <sub>TOOL</sub> [the chest] <sub>OBJECT</sub>	[punch] <sub>GOAL</sub> [with fist] <sub>TOOL</sub> [the stomach] <sub>OBJECT</sub>	[hit] <sub>GOAL</sub> [with elbow] <sub>TOOL</sub> [the back] <sub>OBJECT</sub>	[hit] <sub>GOAL</sub> [with knee] <sub>TOOL</sub> [the back] <sub>OBJECT</sub>	UNINTENTIONAL {NO MATCH}					

In this case visual mapping solved the problem: in fact even for the BSs where the description is not precise, it's easy to say if a video is a good action prototype for it or not.

Given the multilingual nature of the two resources, we could exploit a rich lexical information, i.e. all the verbs in many languages related both to IMAGACT scenes and BabelNet BSs. The connections between BSs and scenes have been automatically established through a Machine Learning algorithm (Gregori et al., 2016).

In order to perform this linking, a dataset of 50 scenes and 57 BabelSynsets (2,850 human judgments in total) have been created<sup>6</sup>. Each ⟨BS, Scene⟩ pair has been evaluated to check if the scene is appropriate in representing the BS. Three annotators compiled the binary judgment table and we reported a Fleiss' kappa inter-rater agreement of 0.74, meaning that at least 2 annotators out of 3 gave the same value for each pair.

IMAGACT data belonging to 17 languages have been exploited to train the algorithm. We used three basic features: the number of verbs connected to the Scene, the number of verbs connected to the BS and the number of verbs that are shared between the Scene and the BS. In each pair ⟨BS, Scene⟩ these features have been calculated for the candidate BS and also for its neighbors, i.e. the other BSs connected through a semantic relation in the BabelNet semantic network.

A Support Vector Machine (SVM) classifier with a RBF kernel have been used to create the model. Table 1 reports Precision, Recall and F-measure.

	Baseline <i>th</i> = 0.04	ML Algorithm 27 features
<b>Pr</b>	0.580	0.833
<b>Re</b>	0.529	0.441
<b>Fm</b>	0.553	0.577

Table 1: Precision, Recall and F-measure of BSs to scenes linking task calculated on the test set for the algorithm and the baseline.

### 3.1.1. Results

Both the resources took an advantage from this linking: IMAGACT gained translation information for languages still not implemented in the Visual Ontology, and BSs referring to action verbs obtained a video representation. In Table 2, the detailed numbers of scenes and BSs connected through this linking are shown.

Table 2: IMAGACT-BabelNet linking results.

IM Scenes linked to BS	773
BS linked to Scenes	517
IM English Verbs related to Scenes	544
BabelNet English Verbs related to BS	1,100

## 3.2. IMAGACT and Praxicon

Similarly to the linking with BabelNet, the IMAGACT scenes have been used to connect the information from Praxicon, given that the definitions of concept in the two

<sup>6</sup>The manually annotated training set is published at <http://bit.ly/2jt2cD4>

resources are too different to obtain a proper and extensive match. In fact, the IMAGACT scenes can work as a visual representation for Praxicon concepts and, at the same time, Praxicon syntax could be used to analytically describe, from a motor point of view, all the low-level actions involved in the execution of more complex ones. Differently from the previous linking, in this case it is a totally manual work, consisting in the analysis of each scene, the determination of the physical actions performed, and the annotation in Praxicon syntax of the motor executions. IMAGACT scenes are specifically created to provide a prototypical representation of a lexicalized action concept: every scene is a reference of at least one action verb. For this reason, such a work of annotation allows to derive some interesting results about the relation between motor and lexical level.

### 3.2.1. Results

The linking with Praxicon is still in progress: the results are partial, but we believe that the integration between linguistic and motor knowledge on action is very relevant both for theoretical analysis and robotic applications. From one side an integrated resource is desirable to carry on deep investigations on the relation between language and action, that is a long debated subject in linguistics and neuroscience (Pustejovsky, 1991; Pulvermüller, 2005; Kemmerer and Gonzalez-Castillo, 2010). Praxicon is also exploited for robotic applications (Vitucci et al., 2016; Tsagarakis et al., 2007) and the integration with a linguistic-oriented resource like IMAGACT can be useful to enhance human-robot interaction through natural language. The scene annotation has been accomplished on 281 IMAGACT scenes (~28% of the total) and we obtained the following results<sup>7</sup>:

- 154 scenes (~55%) have a one-to-one relation with Praxicon Action concepts;
- 64 scenes (~23%) map on more than one Action concept;
- 30 (~11%) are Events but not Actions (in the Praxicon framework);
- 19 scenes (~7%) are Movement but not Actions (in the Praxicon framework);
- 14 scenes (~5%) are unclear.

This data rises some interesting observations about the relation between action verbs and the motor information they express. Consider these two sentences:

1. *John pushes the door;*
2. *John opens the door.*

The verb *to push* (sentence 1) focuses on the performance of the physical action (the *pushing* event) and not on the result, that depends on contextual factors: a closed door will

open by pushing; an open door will close by pushing. Otherwise the verb *to open* (sentence 2) has a specific focus on the result without providing information about the physical action required to reach it: *pushing, smashing, turning the key*, and so on.

Verbs that focus on the action performance (like *to push, to gallop* or *to brush*) reflect some motor features in their semantics. For example the action described in sentence 1 has some motor features that are also encoded in the semantics of *to push*: the application of a force on an object and the outbound direction of the movement. Conversely, verbs that focus on the action result (like *to break, to open* or *to hang*) do not encode specific motor features, given that the result is achieved by performing a set of different physical actions.

This difference mirrors in the annotated scenes: in fact scenes connected to verbs that focus on the performance have a one-to-one relation with Praxicon Action concepts, while scenes connected to verbs that focus on the result usually map on more than one Action concept.

Another thing that emerged from the annotation is the presence of some verbs (like *to drive, to clean* or *to rob*) that predicate complex activities, which are characterized by a high number of physical action that varies a lot depending on the context. In a sentence like *John drives the car*, the activity involves a sequence of actions performed within a loose temporal structure: *turning the steering wheel, pushing the pedals, moving the gearshift*, and so on. The scenes connected to these verbs are considered Events in the Praxicon Framework, and not Actions.

Finally, the scenes that depict a non-voluntary motor activation (like *John falls down*) does not have a goal, so they are not considered Actions, but Movements in Praxicon Framework.

## 4. Conclusions

We introduced the visual mapping methodology that allows resource linking through visual representations. This approach is particularly useful when it's hard to find relations between concepts, as in the representation of actions and events, because it does not force any kind of convergence between senses. For this reason, we feel confident that this methodology could be successfully applied also in other linking tasks involving multimodal resources.

Two case studies have been described: the linking of IMAGACT with BabelNet and Praxicon. In the first case we were dealing with two lexical-semantic resources having huge differences in sense discrimination, and for this reason it was hard to find inter-resource semantic relations. In the case of Praxicon we applied visual mapping to link IMAGACT with a resource of a different type, in which the concepts are motor and not linguistic. This allowed us to derive some preliminary considerations on the relation between linguistic and motor level in action semantics.

## 5. Bibliographical References

Cimiano, P., McCrae, J., Buitelaar, P., and Montiel-Ponsoda, E. (2013). On the role of senses in the ontology-lexicon. In *New trends of research in ontologies and Lexical resources*, pages 43–62. Springer.

<sup>7</sup>At the moment, due to the unfinished state of this task, the inter-annotator agreement have not been calculated.

- Croft, W. and Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge University Press.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
- Christiane Fellbaum, editor. (1998). *WordNet: an electronic lexical database*. MIT Press.
- Gregori, L. and Panunzi, A. (2017). Measuring the italian-english lexical gap for action verbs and its impact on translation. *SENSE 2017*, page 102.
- Gregori, L., Panunzi, A., and Ravelli, A. A. (2016). Linking imagact ontology to babelnet through action videos. *CLiC it*, page 162.
- Kemmerer, D. and Gonzalez-Castillo, J. (2010). The Two-Level Theory of verb meaning: An approach to integrating the semantics of action with the mirror neuron system. *Brain and Language*, 112(1):54–76.
- Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.
- Majid, A., Bowerman, M., van Staden, M., and Boster, J. S. (2007). The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics*, 18(2):133–20.
- Moneglia, M. and Panunzi, A. (2007). Action predicates and the ontology of action across spoken language corpora. the basic issue of the semact project. In *M. Alcántara, T. Declerck, In International Workshop on the Semantic Representation of Spoken Language (SRSL7). Salamanca: Universidad de Salamanca*, pages 51–58.
- Moneglia, M., Brown, S., 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 *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.
- Moneglia, Massimo; Panunzi, A., (2010). *Language, Cognition and Identity. Extension of the Endocentric/Esocentric Typology*, chapter I verbi generali nei corpora di parlato. Un progetto di annotazione semantica cross-linguistica, pages 27–46. Florence University Press.
- Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.
- Pastra, K. and Aloimonos, Y. (2012). The minimalist grammar of action. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367(1585):103–117.
- Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature reviews. Neuroscience*, 6(7):576.
- Pustejovsky, J. (1991). The syntax of event structure. *Cognition*, 41(1):47–81.
- Resnik, P. and Yarowsky, D. (1999). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural language engineering*, 5(2):113–133.
- Tsagarakis, N. G., Metta, G., Sandini, G., Vernon, D., Beira, R., Becchi, F., Righetti, L., Santos-Victor, J., Ijspeert, A. J., Carrozza, M. C., et al. (2007). icub: the design and realization of an open humanoid platform for cognitive and neuroscience research. *Advanced Robotics*, 21(10):1151–1175.
- Vitucci, N., Franchi, A. M., and Gini, G. (2016). Programming a humanoid robot in natural language.
- Wittgenstein, L. (1953). *Philosophical investigations* (gem anscombe, trans.).