# Collecting information for action understanding. The enrichment of the IMAGACT Ontology of Action

Andrea Amelio RAVELLI, Lorenzo GREGORI and Alessandro PANUNZI LABLITA - Università degli Studi di Firenze, Italy

#### Abstract.

This paper presents the status of our work aimed at enriching the IMAGACT Ontology of Action by linking it to other resources. In order to achieve this goal we performed a visual mapping, exploiting the IMAGACT visual component (video scenes that represent physical actions) as the linkage point among resources. By using visual objects, which are free from linguistic constraints and can be interpreted and described from different perspectives, 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 linking obtained by using this technique: an automatic linking between IMAGACT and BabelNet, a multilingual semantic network, and a manual linking between IMAGACT and Praxicon, a conceptual knowledge base of action.

Keywords. ontology linking, IMAGACT, BabelNet, Praxicon,

## 1. Introduction

Action verbs contain the basic information that should be understood in order to make sense of a sentence and that should be processed in instructions given to artificial systems. The difficulty behind action verbs understanding comes out from the evidence that no one to one correspondence can be established between action predicates and action concepts. The same action can be predicated by multiple verbs (e.g. "John *takes/brings/leads* Mary to the restaurant") and, conversely, one verb can extend to multiple and different actions (e.g. "John *takes* the cup from the table", "John *takes/brings* the cup to Mary"). Most of these verbs belong to the class of general verbs, which are characterized by a high ambiguity and high frequency in the use [1]. In these circumstances, senses are often vague and overlapping, their discrimination is not clear, and this is a critical issue for their semantic representation.

The representation is more difficult in a multilingual perspective, given that different languages operate different action space segmentations. It has been observed [2] that even with a fine-grained sense distinction it is not often possible to find an exact match between action concepts lexicalized by verbs in different languages. Moreover, one language could totally lack a lexical representation for a specific concept, whenever there is a lexical gap [3]. These problems deeply affect NLP task dealing with actions and their correct interpretation [4].

This paper reports two linking experiments performed on the IMAGACT Visual Ontology of Action, in order to gather information about actions from several perspective and at different levels: semantic, motoric and visual. Linkings have been led exploiting the visual information of IMAGACT: instead of a classic concept-to-concept mapping we performed a *visual mapping*, that is a concept-to-video linking. This strategy allowed us to connect linguistic resources having different conceptualizations of events.

This work, far from being definitive, could be useful for the future construction of integrated resources on action understanding, to be effectively exploited for both theoretical analysis and computational applications.

## 2. The IMAGACT Visual Ontology of Action

Verbs are the lexical class that is normally responsible for event categorization. Among events, actions (defined as goal-oriented events performed by an intentional agent) play an important role from a linguistic perspective: action verbs are very frequent in spoken language and they are also very ambiguous. Moreover the semantic classification of action verbs is more complex and not equally linear as the one of nouns, so that frequently it's not possible to discriminate a coherent list of word senses.

IMAGACT Visual Ontology of Action<sup>1</sup> [5] 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 form Italian and English spoken corpora. Metaphorical and phraseological usage have been excluded from the annotation process, in order to collect exclusively occurrences of verbs referring to physical actions.

Verbs in IMAGACT are divided into *action types*, according to their semantic variation; each type is linked to one or more video scenes (either 3D animations or filmed video clips), in which a prototypical action is performed. The verbs referring to the same concept are linked to the same scenes, creating an interlinguistic semantic network.

The ontology is in continuus development and, at present, contains 9 fully-mapped languages and 13 that are underway, 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 specifies the lexicon used to express each one in ordinary communication, in all the included languages.

#### 3. Linking resources, sharing knowledge

In order to collect more information, we planned an extensive campaign of enrichment, through the comparison and mutual exchange with other resources.

For this task we applied a visual mapping, a methodology which aims at pointing concepts to a shared visual representation. In fact, a video depicting an event is not subject to any linguistic constraint, and the associated semantic information can be described in various manners. Starting from this observation we used the videos to link concepts of

<sup>1</sup>http://www.imagact.it/

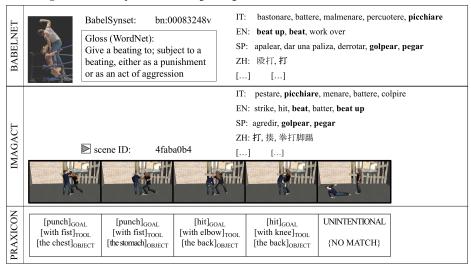


Figure 1. An example of the resulting linking between BabelNet, IMAGACT and Praxicon.

different resources, that express independent event conceptualization according to their own theoretical framework. It follows that the multimodal feature of IMAGACT is a key point for its enrichment and implementation.

Herein we present some current results we obtained by linking IMAGACT with BabelNet and Praxicon. An example of the obtained output can be observed in Figure 1, that shows a *beating* event with the parallel representation in the 3 resources.

## 3.1. IMAGACT and BabelNet

BabelNet<sup>2</sup> [6] is a multilingual semantic network obtained through the automatic mapping of the WordNet thesaurus and the Wikipedia enciclopedia. At present, BabelNet 3.7 contains 284 languages and it is the widest multilingual resources available for semantic disambiguation. Concepts and 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.

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 their 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>&</sup>lt;sup>2</sup>http://www.babelnet.org/

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<sup>3</sup>.

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 on the basis of the number of shared verbal lemmas through an ML algorithm [7].

As a result from this linking, on the one hand, IMAGACT gained translation information for languages still not implemented in the Visual Ontology and, on the other, BSs referring to action verbs obtained a video representation. In Table 1, the detailed numbers of scenes and BSs connected through this linking are shown.

Table 1. 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

Praxicon<sup>4</sup> is an ontology for the representation of action concepts, based on the Minimalistic Grammar of Action [8]. 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 [9].

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

Similarly to the linking with BabelNet, the IMAGACT scenes are used to connect the information of the two resources, given that their definitions of concepts are too different to try a proper and extensive sense matching. In fact, the IMAGACT scenes can work as a visual representation for Praxicon action concepts and, at the same time, Praxicon syntax could be used to analytically describe, from a physical-motoric 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 and the determination of the physical action performed.

<sup>&</sup>lt;sup>3</sup>The measured inter-rater agreement for this task is a Fleiss k of 0.74 with 3 annotators. Annotated dataset is available at http://bit.ly/2jt2cD4

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

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

The scene annotation has been accomplished on 281 IMAGACT scenes ( $\sim$ 28% of the total) and we obtained the following results:

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

IMAGACT scenes are specifically created to provide a prototypical representation of a lexicalized action concept: every scene is a reference of at least an English action verb. This allowed us to derive from these numbers some considerations about the relation between motoric and lexical level.

In Praxicon Events motoric properties does not play a role in the verb meaning, which encodes an abstract result, that is independent from the physical action execution. Examples are verbs like *to drive, to clean* or *to rob*, that encode a complex set of motoric actions by predicating their final result: ~11% of the actions that are commonly referred with the language (in English) belong to this class. Conversely, ~55% of the scenes have a one-to-one mapping with a Praxicon concept, meaning that there is a low distance between motoric and lexical level: we can consider these cases as the ones where the physical execution of an action more deeply affect the verb semantics. Example verbs of this class are *to push, to gallop* or *to brush*. Then, ~23% of retrieved action are at an intermediate level of abstraction: they can be expressed in terms of physical action concepts, but more than one Praxicon concept is involved into a single lexicalized action. Some example verbs are *to break, to open* or *to glue*. Finally, we found that ~7% of events that in English are referred through action verbs are Movement that do not correspond to voluntary actions, like *to fall* or *to drop*.

This work is still in progress, but we believe that the integration between linguistic and motoric 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 [10,11]. Praxicon is also exploited for robotic applications [12,13] and the integration with a linguistic-oriented resource like IMAGACT can be useful to enhance human-robot interaction through natural language.

## 4. Conclusions & Future Works

In this paper we presented the very first steps in the construction of a comprehensive resource for the understanding of actions and their representation in the language systems, built on top of the ontological structure of IMAGACT.

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, 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 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 motoric and not linguistic.

Finally, to extend the information connected to action concepts, we aim to enrich our ontology with the annotation of noun senses and with the predicate argument structures [14], in order to implement semantic selection restriction for the verbs in each action type.

### References

- M. Moneglia & A. Panunzi (2010). I verbi generali nei corpora di parlato. Un progetto di annotazione semantica cross-linguistica. In E. Cresti & I. Korzen (eds), *Language, Cognition and Identity. Extension* of the Endocentric/Esocentric Typology. Firenze: Firenze University Press, 27-46.
- [2] M. Moneglia & A. Panunzi (2007). Action Predicates and the Ontology of Action across Spoken Language Corpora. The Basic Issue of the SEMACT Project. In M. Alcántara Plá & T. Declerck (eds), Proceeding of the International Workshop on the Semantic Representation of Spoken Language. Salamanca: Universidad de Salamanca, 51-58.
- [3] L. Gregori & A. Panunzi (2017). Measuring the Italian-English lexical gap for action verbs and its impact on translation. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations* and their Applications. Valencia: Association for Computational Linguistics, 102-109.
- [4] M. Moneglia (2014). Natural Language Ontology of Action: A Gap with Huge Consequences for Natural Language Understanding and Machine Translation. In Z. Vetulani & J. Mariani (eds), Human Language Technology Challenges for Computer Science and Linguistics, volume 8387 of Lecture Notes in Computer Science. Springer International Publishing, 379-395.
- [5] M. Moneglia, S. Brown, F. Frontini, G. Gagliardi, F. Khan, M. Monachini & A. Panunzi (2014). The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the Representation of Lexical Encoding of Action. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14)*, Reykjavik, Iceland. European Language Resources Association (ELRA), 3425-3432.
- [6] R. Navigli & S. Ponzetto (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence* 193, 217-250.
- [7] L. Gregori, A. Panunzi & A.A. Ravelli (2016). Linking IMAGACT ontology to BabelNet through action videos. In A. Corazza, S. Montemagni & G. Semeraro, *Proceedings of the Third Italian Conference* on Computational Linguistics CLiC-it 2016. 5-6 December 2016, Napoli. Accademia University Press, 162-167.
- [8] K. Pastra, & Y. Aloimonos (2011). The minimalist grammar of action. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 1585, 103-117.
- [9] J. Deng, W. Dong, R. Soecher, L.J. Li, K. Li & L. Fei-Fei (2009). ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Computer Vision and Pattern Recognition (CVPR)*.
- [10] J. Pustejovsky (1991). The syntax of event structure. Cognition 41:1-3, 47-81
- [11] F. Pulvermller (2005). Brain mechanisms linking language and action. *Nature reviews. Neuroscience* **6.7**:576.
- [12] N. Vitucci, A. M. Franchi & G. Gini (2016). Programming a humanoid robot in natural language: an experiment with description logics. *Workshop Simulation in robot programming, SIMPAR 2016*.
- [13] N. G. Tsagarakis, G. Metta, G. Sandini, D. Vernon, R. Beira, F. Becchi, L. Righetti, J. Santos-Victor, A. J. Ijspeert, M. C. Carrozza & D. G. Caldwell (2007). iCub: the design and realization of an open humanoid platform for cognitive and neuroscience research. *Advanced Robotics* 21:10.
- [14] E. Jezek, B. Magnini, A. Feltrcco, A. Bianchini & O. Popescu (2014). A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.