

# From Synsets to Videos: Enriching ItalWordNet Multimodally

Roberto Bartolini, Valeria Quochi, Irene De Felice, Irene Russo, Monica Monachini

Consiglio Nazionale delle Ricerche  
Istituto di Linguistica Computazionale “A. Zampolli”  
name.surname@ilc.cnr.it

## Abstract

The paper describes the multimodal enrichment of ItalWordNet action verbs’ entries by means of an automatic mapping with a conceptual ontology of action types instantiated by video scenes (ImagAct). The two resources present significant differences as well as interesting complementary features, such that a mapping of these two resources can lead to an enrichment of IWN, through the connection between synsets and videos apt to illustrate the meaning described by glosses. Here, we describe an approach inspired by ontology matching methods for the automatic mapping of ImagAct video scenes onto ItalWordNet. The experiments described in the paper are conducted on Italian, but the same methodology can be extended to other languages for which WordNets have been created, since ImagAct is available also for English, Chinese and Spanish. This source of multimodal information can be exploited to design second language learning tools, as well as for language grounding in action recognition in video sources and potentially for robotics.

**Keywords:** action ontology, multimodality, WordNet

## 1. Introduction

Enriching existing lexical resources with multimodal information, both pictures (e.g. ImagNet, Deng et al. (2009)) and videos, is a recent trend in NLP, motivated by practical applications such as second language learning. Dictionaries and lexicographic resources such as WordNet are enriched with multimodal content, because pictures are effective in conveying the meaning of denotative words such as concrete nouns, while for abstract relations (instantiated by prepositional meanings) schematic illustrations can depict several semantic properties. Conveying the meaning of verbs with static representations is not possible; for such cases the use of animations and videos has been proposed (see Stein (1991) cited in Lew (2010)). Short videos depicting basic actions support the user’s need (especially in second language acquisition) to fully understand the range of applicability of verbs i.e. to start with a mental image of an action and from this image find out the L2 verb(s) that can be used to describe that action. More recently, multimodal information has been considered for the possibility to use in a novel way textual information, coupling it with features derived by image and video processing (Bruni et al., 2012).

This paper describes an attempt to automatically enrich ItalWordNet (Roventini et al., 2000) with the ImagAct conceptual ontology of action types instantiated by 3D videos (Moneglia et al., 2012a)<sup>1</sup>. ItalWordNet (IWN henceforth) is a dictionary akin to WordNet in terms of relations among verbs (hyponymy/hyperonymy, troponymy, entailment) and, as its English counterpart, can be used as a lexical ontology for NLP since each synset can be considered as denoting a concept. The ImagAct ontology has been derived annotating occurrences of 600 Italian action verbs highly frequent in spoken corpora; the ontology is structured along 1100 basic action types which we refer to when using an action verb. The two resources present significant

differences in sense splitting: IWN is informed by lexicographic principles while in ImagAct meaning variations for action verbs depend on the way annotators identified difference in body movements performed by human agents interacting with objects and other agents. The mapping between these two resources may thus lead to a multimodal enrichment of IWN, i.e. with ImagAct action types in the form of video scenes mapped on synsets.

## 2. Background and Related Works

Because the integration of existing ontologies is a hot topic especially within the IR and Knowledge management communities, there have been numerous attempts to deal with the problem of ontology merging, alignment, mapping, integration using a variety of methods (e.g. McGuinness et al. (2000), Noy and Musen (2001), Rodríguez and Egenhofer (2003), see also Wache et al. (2001), for a review). A family of approaches views ontologies as graphs and exploits shortest path distance between nodes to assess similarity/proximity and thus propose the mappings (e.g. Cuadros and Rigau (2008)). Given the characteristic of the resources we want to map, however, such methodologies are not applicable to our case.

Another family of approaches tackles the issue of mapping ontologies by means of terminology or lexical matching (cfr. Sánchez et al. (2012)). In this line, Rodríguez and Egenhofer in particular devise an interesting set-theoretical similarity-based method for establishing links between two different ontologies while keeping them autonomous (2003). They implement a matching algorithm that makes use of synonymy sets, distinguishing features of concepts and semantic neighborhood (based on the semantic relation between the concepts) and compute similarity according to the feature-based similarity measure defined by Tversky (1977). Their results prove that synonym-set similarity based on word-matching is a good, though basic, strategy for finding mappings of similar entities across different ontologies. Semantic-neighborhood similarity increases the performance of the system, while including fea-

<sup>1</sup><http://lablita.dit.unifi.it/projects/IMAGACT>

ture matching has a negative effect. Also, while precision is attested around 85%, recall in mapping different ontologies is reported to be low, around 50-55%.

Based on such findings, in the current experiment we adopt a similar set-theory approach and borrow the similarity measure based on the normalization of Tversky's model, which takes into account the differences among sets (see section 3.3. below for details). Similarly to Rodriguez and Egenhofer (2003) synonym set similarity is computed on the basis of word matching (i.e. on the set intersection of the elements -words- that define the entity - wordnet synsets and ImagAct action types). We also take into account the semantic neighborhood by considering hyperonymy relations present in both resources.

### 2.1. ImagAct: a multimodal ontology of action

The ImagAct ontology is composed of basic action types that have been derived bottom-up by annotating high frequency action verbs (approximately 600 lexical entry for each language, Italian and English) extracted from spoken corpora and manually clustering their occurrences on the basis of body movements on objects/agents involved (as in *Fabio rompe lo specchio*, 'Fabio cracks the mirror', and *Il cuoco rompe la noce*, 'the chef breaks the walnut' (see Moneglia et al. (2012a) and Frontini et al. (2012) for details).

Its nodes consist of videos created as 3D animations, each one provided with the sentence that best exemplifies it, according to annotators; each short video represents a particular type of action (e.g. a man taking a glass from a table) and it is related to a list of Italian and English verbs that can be used to describe that action<sup>2</sup>. The 3D animations represent the gist of action in terms of movements and interactions with the object in a pragmatically neutral context. The ontology can be accessed by lemma as well as by scene. Scenes are organized in nine macrocategories (facial expressions, actions referring to the body, movement, modification of the object, deterioration of an object, force on an object, change of location, setting relation among objects, actions in the intersubjective space).

For example, the "modification of an object" macrocategory groups 313 scenes that users can look at<sup>3</sup>. Each video is accompanied by a best example, that shows which Italian verb is usually used to denote the represented action. For instance, the first scene in Figure 1, subtype of the "modification of an object" category, has as best example (BE) the sentence *Fabio suona il pianoforte* ('Fabio plays the piano).

More interestingly, accessing the ontology by scenes sorted by categories, or by single lemmas, allows users to get the list of verbs that refer to each specific action in the languages of the project. To give an example, we may look again at the Italian verb *suonare*, that can describe two action types (fig. 1): the action of playing an instrument as well as that of ringing the bell.

<sup>2</sup>Currently, the ImagAct resource also includes Chinese and Spanish verbs, linked to video scenes.

<sup>3</sup>See <http://imagact.it/imagact/query/gallery.seam>

ImagAct and ItalWordNet ontologies are quite different one from the other. ItalWordNet is modeled on WordNet, one of the best-known lexical resources that contains one of the most complete verbal ontologies, not only in terms of lexical entries, but also in terms of the number of relations among verbs (hyponymy/hypernymy, troponymy, entailment). WordNets have been created for several languages as lexical monolingual databases that group words into sets of synonyms (synsets) and make explicit the various semantic relations among them. In ItalWordNet, every synset thus contains a group of synonymous words or collocations and every word sense appears in only one synset. While originally grounded in cognitive/psychological properties, wordnets are essentially based on lexical meanings of words. As in computational systems (e.g. word sense disambiguation, machine translation, etc.), WordNets are often treated as lexical ontologies: i.e. synsets are taken as denoting concepts, so a comparison between the two resources is meaningful.

The most evident differences between ItalWordNet and ImagAct are the following. First of all, ItalWordNet takes into account the entire lexicon of a language, whereas ImagAct only considers the domain of action verbs. Another great divergence is found in the purposes of the two resources: ImagAct aims to list the different concepts (one or more) which we refer to when using action verbs, whereas WordNet aims to describe all different uses of a verb (including idiomatic or metaphorical expressions).

The two resources, however, also present interesting complementary features: for instance ImagAct does not show semantic relations among verbs, nor it uses definitions/glosses to define actions or action types, while WordNet does; on the other side, WordNet does not distinguish between primary and marked senses, often collapsing concrete uses with metaphorical or idiomatic ones<sup>4</sup>.

Furthermore, ItalWordNet defines horizontal relations among senses (synsets) with glosses, while ImagAct uses scenes to represent the event type which different verbs can refer to in similar contexts (equivalent verb classes).

### 3. Mapping ImagAct on ItalWordNet

Given the potential complementarity of the two resources, mapping them could lead to a reciprocal enrichment: for example, in case of perfect matching between an action type and a synset, ImagAct videos might be enriched by IWN glosses, and IWN glosses could be more intuitively understood if visually represented.

In this work, we experiment on the possibility of establishing an automatic mapping between the ImagAct ontology and WordNet resources and focus especially in the potential enrichment of ItalWordNet with video scenes.

In the following, we describe the experiment conducted in this direction, present an evaluation of the method and

<sup>4</sup>For instance, consider the first sense for the verb *to fill* that is both concrete and metaphorical:

- fill, fill up, make full (make full, also in a metaphorical sense) "fill a container"; "fill the child with pride";

|   |   |
|---|---|
|  | <p><i>Suonare</i> Type 1 (Scene: bbc50559)<br/> Best Example: <i>Fabio suona il pianoforte</i><br/> English "to play"<br/> Chinese "弹 tǎn"<br/> Spanish "tocar"</p>                   |
|  | <p><i>Suonare</i> Type 2 (Scene: 4b8bcda1)<br/> Best Example: <i>Marta suona il campanello</i><br/> English "to ring, to sound"<br/> Chinese "按 àn"<br/> Spanish "tocar (timbre)"</p> |

Figure 1: Representation of two "play" scenes

finally make some general considerations about the outcomes.

### 3.1. Basic Assumptions

In the ImagAct ontology basic action types are represented by scenes (usually described by means of short 3D videos) and each scene is associated to a set of verbs (types). Scenes can thus be seen as sets of "(locally) equivalent" verbs types, which all together may be taken as representative of a concept.

For the purpose of the current mapping, the scenes and their sets of "locally equivalent" verb lemmas are then equiparated to ItalWordNet synsets and, similarly, ImagAct verb types can be seen as akin to lexical word senses.

Relying on this assumption, our working hypothesis is that we can automatically establish correspondences between IWN verbal synsets and ImagAct basic action types by measuring the semantic proximity between video scenes and synsets in terms of overlap between equivalent verbs (lemmas) in ImagAct and synonyms and hyperonyms in IWN.

For instance, the Italian verb *pelare* has only one ImagAct action type, denoting the action of 'skinning vegetables and fruits' and is associated to the scene id:a8b7753e together with only one equivalent verb *sbucciare*. The same verb has five senses in ItalWordNet, identified by the following synsets and hyperonym synsets:

1. SYNSET: (pelare [1]), 'to skin' (animals)  
HYPERONYM: (privare [1], togliere [2]), 'to take away'
2. SYNSET: (pelare [2], rapare [1]), 'to crop, to cut (hair)'  
HYPERONYM:( radere [1]), 'to shave'
3. SYNSET: (pelare [3], spennare [1], spiumare [1]) 'to deplume'  
HYPERONYM:(strappare [1]), 'to tear off'
4. SYNSET: (pelare [4], sbucciare [1]), 'to skin (vegetables and fruits)'  
HYPERONYM: (privare [1] togliere [2]), 'to take away'
5. SYNSET: (pelare [5], tosare [4]), 'to rip off'  
HYPERONYM: (spogliare [1]), 'to strip (belongings)'

In this case, the correct mapping for ImagAct *pelare* is IWN Sense 4, that exactly refers to the action of skinning vegetables and fruits.

Therefore, we would like the algorithm (cfr. §3.3.) to be able to choose IWN Sense 4 as the best candidate synset for the mapping, because it has *sbucciare* in its synset, thus it matches better than other candidates with the set of ImagAct equivalent verbs.

### 3.2. Data

From the ImagAct resource database we collect, for every video scene Id, all its related Italian verb lemmas together with the relation they bare with the scene (i.e PROTO or INST)<sup>5</sup>. Each verb in the Imagact project, in fact, was (manually) annotated as either prototypical (PROTO) or as an instance (INST) relative to the video scene, where INST means that the verb denotes a more general action than the one represented in the scene<sup>6</sup>

The Imagact dataset used for the mapping consists of 1120 video scenes with a total of 1100 associated Italian verb types (500 lemmas, with an average of 2.4 verb lemmas per scene<sup>7</sup>).

Concerning ItalWordNet we consider as relevant information: verbal synsets, verb senses, hyponymy and hyperonymy relations. Altogether, the ItalWordnet database (hosted at CNR-ILC) contains 8903 verbal synsets, 14086 verb senses (8121 lemmas, with an average of 1.1 verb lemmas per synset) that are potential candidates for the mapping.

### 3.3. The Mapping Algorithm

In this section we describe the algorithm implemented for the mapping of ImagAct to ItalWordNet. For the sake of clarity, let:

- $\Omega$  be the set of ImagAct video scene identified by their sceneId  $\omega$ ;

<sup>5</sup>As the ontology is constructed bottom-up it is in fact not (yet) formalised, and data are stored in DB format at LABLITA, University of Florence.

<sup>6</sup>Notice that here the distance between the INSTance verb and the target scene is not known.

<sup>7</sup>The ImagAct final database contains about 600 verb lemmas; however, at the time we ran the experiment, only 500 of them were considered validated.

- $V_\omega = \{v_1, v_2 \dots v_n\}$  be the set of verb element  $v_i, \forall i = 1 \dots n$  of a scene with sceneId  $\omega$ ;
- $W$  be the set of all word senses, with part of speech Verb, and  $S$  be the set of Synsets of ItalWordNet, considered extensionally, that is taking into account the elements  $s_j$  that compose it;
- $\Gamma : W \rightarrow S$  be defined by  $\Gamma(w) \stackrel{def}{=} s$  iff  $w \in s$ . (Notice that  $\Gamma$  is well defined since every word sense belongs to a single synset).

Thus, for each ImagAct video scene (id) we consider the set  $V$  of verbs associated with the scene and the specific relation with the video scene as a feature. For each verb associated to the video scene, we search ItalWordNet for the list of its possible senses and for each of the senses we retrieve the synset ID. This way, each verb of the video scene will be associated to a list of synset IDs.

For each synset ID, we then retrieve the whole set of verb lemmas it contains in ItalwordNet (i.e we consider the original synset), and, for each synset that contains a PROTO verb of the scene, an extended set is created by including the set of its verbal lemmas and the set of verbal lemmas of all its hyperonyms. Both sets are then considered together as the set (of sets)  $\Sigma$ .

Once the feature set for all ImagAct scenes and ItalWordNet synsets are retrieved as described above, similarity between the sets has to be calculated in order to propose the mappings.

In the present work, the Jaccard index (Jaccard, 1912) and the Tversky ratio model (Tversky, 1977) are the similarity measures used to assign a score to each set in  $\Sigma$  relative to the video scene  $V$ . We also introduce a weighting mechanism for cases where synsets contain only one lemma (not infrequent in IWN). If the scene has only one lemma, the score is positively weighted for the synset with the lower ID number (which corresponds to the more basic/concrete sense for that verbal lemma<sup>8</sup>). As a result of this step, we obtain a ranking of possible matches for each video scene/verb type.

Finally, the highest score is taken for selecting the best candidate(s) for matching. That is, the synset(s) that receive the higher similarity score is(are) proposed as the best mapping candidate(s) for the ImagAct video scene/verb type. Indeed, since more than one synsets may receive equal score, the system can propose more than one mapping synsets. Formally, the algorithm implemented for the mapping of ImagAct video scenes on IWN is the following:

1.  $\forall \omega \in \Omega$  consider the elements of scene  $V_\omega$
2.  $\forall v_i \in V_\omega$  with  $i = 1 \dots n$  consider its  $m$  sense  $v_i^1, v_i^2 \dots v_i^m$  then
3.  $\forall j \in \{1 \dots m\}$  build the set of synset  $\Sigma_\omega$  containing  $\Gamma(v_i^j)$ ;<sup>9</sup> therefore  $\Sigma_\omega = \{\Gamma_{v_i^j} | v_i^j \xrightarrow{\Gamma} \Gamma(v_i^j) \equiv \Gamma_{v_i^j} \text{ and } v_i \in V_\omega\}$ ;

<sup>8</sup>This is systematic in ItalWordNet: synsets referring to more basic/concrete concepts are assigned lower IDs.

<sup>9</sup>Note that  $m$  is not a constant and it depends on the verb  $v_i$ .

4. if the type of  $v_i$  is PROTO then we extend  $\Sigma_\omega$  to  $\Sigma_\omega^+$  to take into account the hyperonymy information: that is  $\forall j \in \{1 \dots m\}$  we add to  $\Sigma_\omega \forall k = 1 \dots p$  the sets  $\Gamma_{v_i^j} \cup \Psi_{v_i^j}^k$ .<sup>10</sup>
5. Let  $\lambda$  the similarity measures used then calculate  $\lambda(\omega, \sigma) \forall \omega, \sigma : \omega \in \Omega$  and  $\sigma \in \Sigma_\omega^+$ . Let  $\Lambda_\omega = \{\sigma | \lambda(\omega, \sigma)\}$
6. Then build the mapping  $M : \Omega \rightarrow S$  such that  $M(\omega) = \sigma^*$  iff  $\sigma^* \in \Lambda_\omega$  and  $\forall \sigma \in \Lambda_\omega : \lambda(\omega, \sigma) \leq \lambda(\omega, \sigma^*)$

### 3.4. Experiment

For the experiment we implemented two versions of the algorithm in section 3.3., which differ in the similarity index used:

**Jaccard version:** similarity is calculated in a geometrical space by using the plain Jaccard measure ((Jaccard, 1912)):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

**Tversky version:** the mappings are established by calculating the similarity between synsets and scenes according to a set-theoretical approach, as in (Tversky, 1977) and (Rodríguez and Egenhofer, 2003).

$$T(A, B) = \frac{|A \cap B|}{|A \cup B| + \alpha|A \setminus B| + (1 - \alpha)|B \setminus A|}$$

with  $0 \leq \alpha \leq 1$

here  $\alpha$  is a parameter that weights the influence of the differences of the two sets being compared (i.e. of the elements of a set that is absent in the other set) for the final mapping.

The Jaccard version is used as a baseline against which we assess the Tversky one. The methods are applied to the Italian part of the ImagAct Ontology which was available and stable in October 2013 and on the ItalWordNet resource<sup>11</sup>.

## 4. Evaluation and Discussion

The experiment conducted focuses on the mapping of IWN and the Italian verbs associated to the action types in the ImagAct ontology. The results have been evaluated in terms of Precision, Recall and F-measure against the manually annotated gold standard. In this section we briefly describe the gold standard, then we present and discuss the evaluation and some qualitative analysis of the results.

We will use as notation  $\Gamma_{v_i^j} \equiv \Gamma(v_i^j)$  to denote the synset associated to wordSense  $v_i^j$  and  $\{\Psi_{v_i^j}^k\}_{k=1 \dots p}$  to denote all the  $p$  hypernyms of the synset  $\Gamma_{v_i^j}$ ; similarly to  $m$  that depends on  $i$ , also  $k$  is a function of  $m$ .

<sup>10</sup>This is done to take into account the available information about neighbourhood of the concepts. The intuition here is that when a synset contains a PROTO verb of a scene and its hyperonym synset contains an INST verb of the same scene, it is more likely that the two concepts are similar.

<sup>11</sup>ItalWordNet is available on Data Hub (<http://datahub.io/dataset/iwn>), while the ImagAct ontology is browsable from a web interface from <http://imagact.it/imagact/query/dictionary.seam>

## 4.1. The gold standard

As gold standard mapping for evaluation, we use a revised and corrected version of a manual mapping of ImagAct verb action types onto IWN synsets, which was prepared in the context of a previous study. The mapping was performed by comparing and manually mapping Imagact verb types onto one or more IWN synsets, both looking at the video scenes and considering the best example(s) associated with each action type (see (Moneglia et al., 2012b), (Frontini et al., 2012) for details).

This gold standard consists of 260 Italian verb lemmas from ImagAct, corresponding to 358 action types, which map onto a total of 343 IWN synsets.

During goldstandard creation 3 possible matching situations were observed: a) cases where a perfect 1:1 match could be identified (the majority of cases), b) and c) cases where an “imperfect” match was found, i.e. where either the verb type or the synset match more than one of its counterparts. The next paragraphs briefly exemplify these cases. Here, AT refers to (verb) Action Types in ImagAct and Syn to synsets in IWN<sup>12</sup>.

- a) **Perfect 1:1 match (AT=Syn)**: in the majority of cases we found a perfect correspondence with one IWN synset, as for *nuotare*, ‘to swim’, in example (1).

- (1) ImagAct\_Action\_Type\_1:  
BE: *Matteo nuota nell’acqua*  
‘Matthew swims in the water’  $\implies$   
IWN\_Sense1:  
Gloss: *muoversi sulla superficie dell’acqua eseguendo movimenti coordinati delle braccia e delle gambe*  
‘to move on the surface of water moving arms and legs in a coordinated way’

- b) **Imperfect 1:n match (AT=Syn+Syn)**: in some cases one action type subsumes more than one synset. As a consequence, we consider this as an imperfect match between one action type and two or, rarely, three synsets, as for *urlare*, ‘to shout’, in example (2).

- (2) ImagAct\_Action\_Type\_1:  
BE: *Fabio urla*  
‘Fabio shouts’  
 $\implies$  ItalWordnet\_Sense\_2:  
Gloss: *parlare a voce troppo alta e in modo sguaiato,*  
‘to talk too loud’  
  
 $\implies$  ItalWordnet\_Sense\_3:  
Gloss: *parlare con tono di voce molto alto, udibile a distanza,*  
‘to talk in a loud voice’.

- c) **Imperfect n:1 match (AT+AT=Syn)**: a few cases were found where 2,3, or 4 action types correspond

to a single synset, i.e. the synset subsumes more than one action type. This is the case of *accostare*, ‘to put (close)’, as in example (3) below.

- (3) ImagAct\_Action\_Type\_1:  
BE: *Fabio accosta il suo viso al viso di Cristina*  
‘Fabio puts his face close to Cristina’s face’ &  
ImagAct\_Action\_Type\_2:  
BE: *Fabio accosta il tavolo alla parete*  
‘Fabio puts the table close to the wall’  $\implies$   
ItalWordnet\_Sense\_1:  
Gloss: *mettere una cosa vicino a più vicino ad un’altra*  
‘to put something close to another thing’

The chart in fig.2 summarises the distribution of the mappings in the gold standard described above.

## 4.2. Evaluation results

The IWN - ImagAct mappings resulting from the experiment have been intrinsically evaluated against the gold standard described in 4.1., in terms of precision, recall,  $F_1$ , and  $F_{0.5}$ <sup>13</sup>). Here, these measures are adapted to the set-theoretical approach followed and defined on the basis of the cardinality of the sets of gold-synsets and retrieved synsets per verb-type.

As the goldstandard was created mapping verb action types (not scenes) to IWN synsets, we automatically assess the performance of the algorithm on the same task of mapping verb types onto synsets. In order to do so, we project the similarity scores obtained at scene level onto the verb type related to the scene and promote as good mapping the synset which receives the highest score which also contains the verb lemma corresponding to the type in Imagact.

Figure (3 illustrates some examples of mapping for the verb *chiudere* (‘to close’). For each scene, we report similarity scores only for the first two candidate synsets (although according to the current procedure, the system would propose the one(s) with the highest score).

The first scene (d846ce14) is pointed at by two ImagAct verbs, *chiudere* (action type 744, ‘to lock’) and *rinchiudere* (action type 1297, ‘to lock’); it shows a prison-guard locking a detainee. The highest similarity score (1.02) was given to synset {32063}, that contains exactly the same verb lemmas (senses {chiudere-3, rinchiudere-1}) that in ImagAct refer to scene d846ce14. Furthermore, this scene well represents the meaning of the gloss of the synset (‘to lock in a delimited place [...]’) and of the corresponding example, referring to birds being locked in a cage. Thus, in this case, a perfect 1:1 match is correctly found.

The second scene (9d7c36a1), which refers to the action of closing parts of the body (hands, mouth, eyes etc.), in

<sup>13</sup> $F_{0.5}$ -measure weights precision higher than recall (for  $\beta=0.5$ ).

Because recall is expected to be low and especially because in this task it is more important that the automatically established mapping are correct than that the system is able to retrieve all manual mappings, we consider  $F_{0.5}$  as a good indicator of the overall performance.

<sup>12</sup>For the sake of exemplification we will use fake ids of action types and synsets.

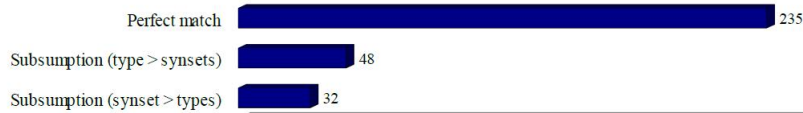


Figure 2: IWN-ImagAct mapping




| Scene ID | Scene Representation  | ImagAct                                      | ItalWordNet   | Similarity Score    |
|----------|---|--|---|---------------------|
| d846ce14 |  | d846ce14<br>{chiudere-744, rinchiudere-1297} | 32063 {chiudere-3, rinchiudere-1}<br>32061 {rinchiudere-3, rinserrare-2} +<br>{chiudere-3, rinchiudere-1} | 1.02<br>0.97561     |
| 9d7c36a1 |  | 9d7c36a1<br>{chiudere-743}                   | 36869 {chiudere-1, serrare-2}<br>32063 {chiudere-3, rinchiudere-1}  | 1.47238<br>0.952381 |
| ee7c4403 |  | ee7c4403<br>{chiudere-912}                   | 36869 {chiudere-1, serrare-2}<br>32063 {chiudere-3, rinchiudere-1}  | 1.47238<br>0.952381 |

Figure 3: Examples of ImagAct-ItalWordNet mapping (verb *chiudere*, 'to close').

Italian can be denoted only by *chiudere* (action type 743, 'to close'). In this case, the algorithm correctly assigns the highest similarity score (1.47238) to synset {36869}, that contains the senses {chiudere-1, serrare-2}. It is worth noting here that this synset also gained exactly the same similarity score with scene ee7c4403, which is denoted too only by *chiudere* (type 912, denoting closing doors, windows etc.). For these two scenes, synset {36869} is chosen as the best candidate by the algorithm for two reasons:

- there are no synsets that contain only the verb *chiudere*;
- when two or more synsets containing the same number of verbs are retrieved (e.g. 36869: {chiudere-1, serrare-2}; 32063: {chiudere-3, rinchiudere-1}; 33596: {chiudere-7, ostruire-1}), preference is given to the one in which the target verb lemma (in this case *chiudere*) is found.

The gloss that defines synset {36869} ('to close something, so that it will not open'), in fact, could perfectly be represented both by scene 9d7c36a1 and by scene ee7c4403. This is also demonstrated by the fact that in ItalWordNet, for this synset, three examples are provided, two of which explicitly refer to body parts and to doors. In this case, therefore, a n:1 match was correctly established.

In the evaluation, we compare the results obtained by running the two versions of the algorithm (in section 3.3.), where the Jaccard version is taken as the baseline system and used to assess the best configuration (i.e. the best  $\alpha$ ) for the Tversky version.

Table 4.2. presents the performance results of the baseline, while Table 2 reports the scores obtained applying the Tversky method at various values for  $\alpha$ .

Interestingly, from these results we note that with  $\alpha$  equal to 0.5, which makes the formula equivalent to Jaccard, we obtain the same results as in the best scenario for experiment 1.

|  | R    | P    | F0.5 | F1   |
|--|------|------|------|------|
|  | 0.59 | 0.64 | 0.63 | 0.61 |

Table 1: Performance of the baseline system: the Jaccard version

| $\alpha$ | R    | P    | F0.5 | F1   |
|----------|------|------|------|------|
| 0.15     | 0.60 | 0.62 | 0.62 | 0.61 |
| 0.35     | 0.60 | 0.64 | 0.63 | 0.62 |
| 0.50**   | 0.59 | 0.64 | 0.63 | 0.61 |
| 0.70     | 0.60 | 0.66 | 0.65 | 0.63 |
| 0.75     | 0.60 | 0.65 | 0.64 | 0.62 |
| 0.80     | 0.59 | 0.66 | 0.64 | 0.62 |
| 0.88     | 0.60 | 0.68 | 0.66 | 0.64 |
| 0.90     | 0.60 | 0.68 | 0.66 | 0.64 |
| 0.95     | 0.61 | 0.71 | 0.69 | 0.66 |
| 1        | 0.61 | 0.69 | 0.67 | 0.65 |

Table 2: Performance of the Tversky version, at different  $\alpha$  values

\*\*equivalent to Jaccard index

With  $\alpha$  close to one, that is giving negative influence to the set differences, instead, the precision increases significantly. Also worth of notice is that recall, though not particularly high, also increases at higher values for  $\alpha$ . Given the results of previous works which show that recall stays low when mapping different and independent ontologies (Rodríguez and Egenhofer, 2003), the low recall obtained is not surprising, given the differences between the two ontologies.

Another explanation of the low recall, is that the evaluation scores reported above are calculated on the set intersection of synset mappings; thus, in cases where the gold-standard indicates more than one corresponding synsets for a given verb type and the system only finds one correct mapping,

recall is greatly affected. In fact, if we run the evaluation considering gold-standard verb-type to synset(s) mapping as singletons, at  $\alpha = 1$ , we obtain a recall score of 0.73.

Finally, in order to assess the actual quality and acceptability of the automatic mapping, we run a qualitative analysis of the false positives generated by the system at  $\alpha = 0.95$ , to check how (in)acceptable the mapped synsets really are relative to the scene.

Such analysis reveals that in 24 cases (21.43%) the automatic mapping proposed is in fact acceptable. This mostly happens when two synsets are almost identical both in meaning and for the verbs they contain, as in the following example:

- (4) Scene 64fa01f7;  
verb *verniciare*  
'to varnish, paint'  
BE: *L'imbianchino vernicia*  
'The painter paints'

**Expected** synset: {34031} (verniciare[1])  
'to varnish'

Gloss: 'cover with a layer of paint, a wall, a cabinet, a fixture, etc..'

**Actual** result: Synset32367 (colorare[3], pitturare[2], verniciare[2])

'to color', 'to paint', 'to varnish'

Gloss: 'paint using colored paints and varnishes'.

## 5. Conclusions

We presented in this paper an ontology mapping experiment with the goal of automatically enriching ItalWordNet with a multimodal ontology of action types (ImagAct), so that synsets denoting concrete actions would be linked to 3D video scenes that further exemplify their meaning<sup>14</sup>.

Given the structural and design differences of the two resources, an automatic mapping is per se a challenge.

The experiment described here implemented an algorithm inspired by Rodríguez and Egenhofer (2003) based on set-theory and feature-based similarity assessment, which proved particularly interesting for the mapping of different and independent ontologies and especially fit for lexical resources, as it is primarily based on word matching. Our results are in line with their findings in terms of performance, which provides an indication that our automatic mapping is fairly reliable. Also, the results seem to prove that, at least for wordnet-like lexical resources, differences in the synonym sets are relevant for assessing the proximity or distance of concepts. Indeed, the Tversky version which almost maximises the weight of such differences obtained the best results.

<sup>14</sup>As the ImagAct ontology is still under formalisation and IRP and exploitation issues are still to be decided, the full mapping is not yet available. It is however our intention to publish it in RDF format as Linguistic Linked Data as soon as possible. A demo with a sample of the mapping is available at <http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=924/vers=ing>

Although the experiment has been tested only on Italian, the same method/system can be easily applied to English, for which have both the Imagact database and WordNet.

As future work, we plan to use the same method mapping the English part of ImagAct onto PWN, and experiment with exploiting the interlingual links between the two wordnets to improve the precision of the mapping.

## 6. Acknowledgements

This work has been carried out within the context of the ImagAct project. The ImagAct project has been funded in Italy within the PAR/FAS program of the Tuscan Region and it was undertaken by the University of Florence, ILC-CNR, Pisa, and the University of Siena.

## 7. References

- Bruni, Elia, Uijlings, Jasper R. R., Baroni, Marco, and Sebe, Nicu. (2012). Distributional semantics with eyes: using image analysis to improve computational representations of word meaning. In Babaguchi, Noboru, Aizawa, Kiyoharu, Smith, John R., Satoh, Shin'ichi, Plagemann, Thomas, Hua, Xian-Sheng, and Yan, Rong, editors, *ACM Multimedia*, pages 1219–1228. ACM.
- Cuadros, Montse and Rigau, German. (2008). Knownet: Building a large net of knowledge from the web. In Scott, Donia and Uszkoreit, Hans, editors, *COLING*, pages 161–168.
- 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*.
- Frontini, Francesca, de Felice, Irene, Khan, Fahad, Russo, Irene, Monachini, Monica, Gagliardi, Gloria, and Panunzi, Alessandro. (2012). Verb interpretation for basic action types: annotation, ontology induction and creation of prototypical scenes. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 69–80, Mumbai, India, December. The COLING 2012 Organizing Committee.
- Jaccard, Paul. (1912). The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50.
- Lew, Robert. (2010). New ways of indicating meaning in electronic dictionaries: hope or hype? In Zhang, Yihua, editor, *Learner's Lexicography and Second Language Teaching*, pages 387–404. Shanghai Foreign Language Education Press, Shanghai.
- McGuinness, Deborah L., Fikes, Richard, Rice, James, and Wilder, Steve. (2000). An environment for merging and testing large ontologies. In Cohn, A. G., Giunchiglia, F., and Selman, B., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000)*, pages 483–493. Morgan Kaufmann.
- Moneglia, Massimo, Gagliardi, Gloria, Panunzi, Alessandro, Frontini, Francesca, Russo, Irene, and Monachini, Monica. (2012a). Imagact: Deriving an action ontology from spoken corpora. In Bunt, Harry C., editor, *Proceedings of the Eighth Joint ACL - ISO Workshop on Interoperable Semantic Annotation (isa-8)*, pages 42–47.
- Moneglia, Massimo, Panunzi, Alessandro, Gagliardi, Gloria, Monachini, Monica, Russo, Irene, and Frontini,

- Francesca. (2012b). Mapping a corpus-induced ontology of action verbs on italwordnet. In Fellbaum, Christiane D. and Vossen, Piek, editors, *Proceedings of the 6th Global Wordnet Conference*, pages 219–226.
- Noy, Natalya F. and Musen, Mark A. (2001). Anchor-prompt: Using non-local context for semantic matching. In *Proceedings of the Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 63–70.
- Rodríguez, M. Andrea and Egenhofer, Max J. (2003). Determining semantic similarity among entity classes from different ontologies. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 15:442–456.
- Roventini, Adriana, Alonge, Antonietta, Calzolari, Nicoletta, Magnini, Bernardo, and Bertagna, Francesca. (2000). Italwordnet: a large semantic database for italian. In *LREC*. European Language Resources Association.
- Sánchez, David, Solé-Ribalta, Albert, Batet, Montserrat, and Serratosa, Francesc. (2012). Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. *Journal of Biomedical Informatics*, 45(1):141–155.
- Stein, Gabriele. (1991). Illustrations in dictionaries. *International Journal of Lexicography*, 4(2):99–127.
- Tversky, Amos. (1977). Features of similarity. *Psychological Review*, 84:327–352.
- Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information - a survey of existing approaches. pages 108–117.