La definizione delle relazioni intra- e interlinguistiche nella costruzione dell'ontologia IMAGACT

Gloria Gagliardi
Università degli Studi di Firenze
gloria.gagliardi@unifi.it

Abstract

IMAGACT è un'ontologia interlinguistica che rende esplicito il *range* di variazione pragmatica associata ai predicati azionali a media ed alta frequenza in italiano ed inglese. Le classi di azione che rappresentano le entità di riferimento dei concetti linguistici, indotte da *corpora* di parlato da linguisti madrelingua, sono rappresentate in tale risorsa lessicale nella forma di scene prototipiche (Rosch 1978). Tale metodologia sfrutta la capacità dell'utente di trovare somiglianze tra immagini diverse indipendentemente dal linguaggio, sostituendo alla tradizionale definizione semantica, spesso sottodeterminata e linguo-specifica, il riconoscimento e l'identificazione dei tipi azionali. L'articolo illustra i criteri generali che hanno ispirato il *mapping* inter-/intra- linguistico dei dati derivati da *corpora* per la formazione dell'ontologia, le questioni di natura teorica e tecnica poste dalla costruzione della risorsa e le soluzioni adottate. Vengono descritte le tipologie e la natura delle relazioni tra le entità del database nella sua versione 1.0, e le modalità generali con cui i materiali linguistici annotati sono stati organizzati in una struttura dati coerente.

Keywords: ontologia; verbi di azione; relazioni interlinguistiche

1 Introduzione

I verbi d'azione veicolano informazioni essenziali per la corretta interpretazione delle frasi e quindi per la comprensione del linguaggio. Le relazioni che strutturano questa parte di lessico sono tuttavia molto complesse, in quanto non è possibile stabilire una corrispondenza biunivoca tra predicati ed eventi (Moneglia & Panunzi 2010). I verbi d'azione più frequenti nella comunicazione quotidiana sono infatti "generali", ovvero possono essere applicati in modo produttivo a classi di azioni pragmaticamente e cognitivamente diverse, come mostra la variazione pragmatica del verbo italiano.

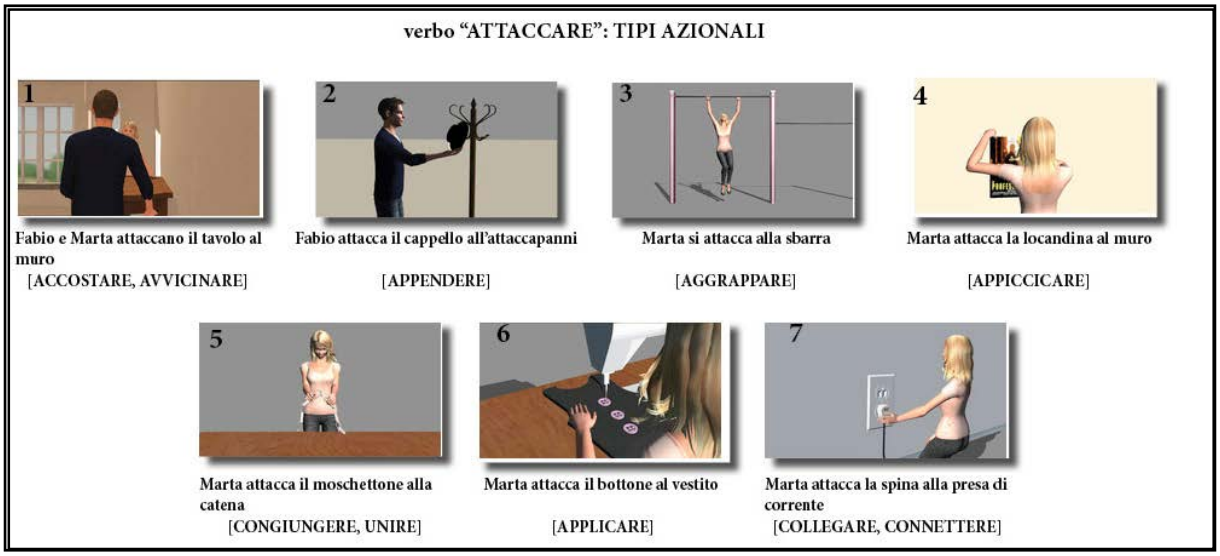

Fig. 1: Variazione pragmatica del lemma italiano *attaccare*.

Tale variazione corrisponde alla competenza semantica referenziale dei parlanti, ed è quindi un dato essenziale per la modellizzazione dell'informazione lessicale. Tuttavia, essa è solo saltuariamente censita dai dizionari tradizionali e dalle più note ontologie e risorse computazionali, come ad esempio Wordnet (Fellbaum 1998) e Verbnet (Kipper-Schuler 2005).

Il progetto IMAGACT contribuisce a superare questa lacuna attraverso la realizzazione di un'ontologia interlinguistica dell'azione che esplicita lo spettro di variazione pragmatica associata ai predicati in italiano e in inglese (Moneglia *et al.* 2012). Le classi di azioni che identificano il riferimento di ogni verbo sono state individuate a partire dall'annotazione di grandi *corpora* rappresentativi dell'uso linguistico parlato spontaneo, e quindi associate attraverso una procedura di *mapping* inter-/intra- linguistico ad una serie di scene prototipiche, in grado di elicitare nell'utente la comprensione della classe di eventi rappresentata (Rosch 1978).

La metodologia di induzione delle classi di azioni e l'utilizzo di prototipi in sostituzione delle definizioni per la rappresentazione del riferimento, due tra gli aspetti più innovativi di IMAGACT, hanno però sollevato alcune questioni di strutturazione dell'informazione nella fase di formazione dell'ontologia (par. 3.1).

Verranno descritti, a partire da alcuni *case study*, i problemi e le soluzioni adottate per la costruzione della risorsa, ovvero le modalità secondo le quali il materiale annotato è stato organizzato all'interno di una struttura dati coerente, che, pur mantenendosi aderente all'intuizione dei parlanti madrelingua, risulti di facile consultazione per l'utente finale.

2 Induzione delle Classi Azionali da Corpus

Una strategia efficace per apprezzare l'uso dei verbi *action-oriented* è la diretta osservazione delle loro occorrenze nel parlato spontaneo, in cui il riferimento all'azione è decisivo. In IMAGACT è stata dunque adottata una procedura di tipo *bottom-up*; le classi di azioni sono state indotte da risorse linguistiche di parlato disponibili, su licenza, per scopi scientifici:

- *corpus* Inglese: una selezione del British National Corpus (BNC) di circa 2 milioni di parole;
- *corpus* Italiano: una collezione di risorse di parlato in lingua italiana (LABLITA corpus, LIP, CLIPS) per un totale di 1,6 milioni di parole (Moneglia *in press*; Gagliardi 2014).
- I materiali linguistici sono stati sottoposti ad una articolata procedura di annotazione (Moneglia *et al.* 2012; Frontini *et al.* 2012); i dati risultanti consistono di:
- due elenchi di verbi, uno per l'italiano e uno per l'inglese;
- per ogni verbo, una serie di "tipi" azionali, ovvero le classi di azioni fisiche tra loro tipologicamente e cognitivamente diverse che rientrano nell'estensione del predicato (fig. 1);
- per ogni tipo azionale, uno o più *Best Example*, cioè le istanze più rappresentative di tutte le strutture tematiche e delle proprietà aspettuali individuate;
- per ogni *Best Example*, l'insieme delle occorrenze che costituiscono la variazione del verbo nel *corpus*, standardizzate da linguisti madrelingua in frasi semplici ed annotate a vari livelli.

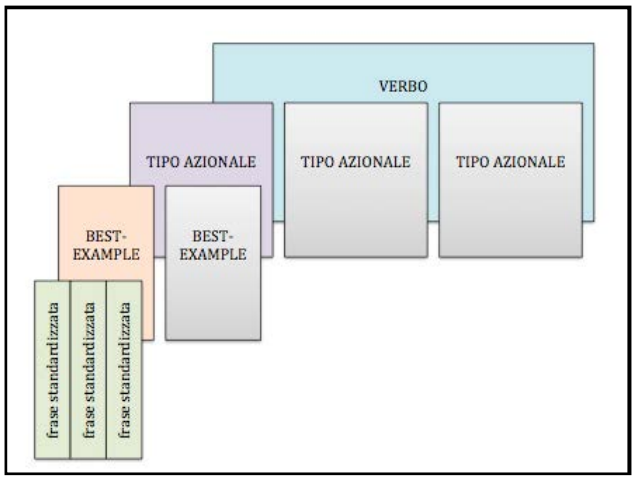

Fig. 2: Risultati della procedura di annotazione IMAGACT.

L'ontologia interlinguistica è stata costruita mediante il ricongiungimento dei tipi azionali in un'unica galleria di scene prototipali. Il requisito generale che ha ispirato la formazione della risorsa è che l'immagine standard associata a ciascun tipo sia facilmente riconoscibile e garantisca la corretta individuazione del concetto azionale, indipendentemente dalla lingua e dalla cultura di origine dell'utente.

3 Mapping

3.1 Criteri generali

La costruzione di un'ontologia coerente dal punto di vista linguistico e formale a partire dai materiali estratti da *corpora* ha rappresentato una sfida molto impegnativa, sia a livello pratico, considerata l'enorme mole di dati da riconciliare in una struttura unitaria, che a livello teorico, data la novità della metodologia di induzione delle classi azionali.

I dati in *input* hanno influenzato fortemente la forma e la struttura del database e la concezione stessa della procedura di *mapping* inter-/intra- linguistico. La tipizzazione della variazione è infatti condizionata dalla semantica del verbo in oggetto: il senso del lemma, operando come "punto di vista" sulle categorie azionali, ha determinato la granularità dell'annotazione, anche a parità di eventi predicabili. Si considerino ad esempio le variazioni dei lemmi *attaccare* ed *appendere* riportate in fig. 3.

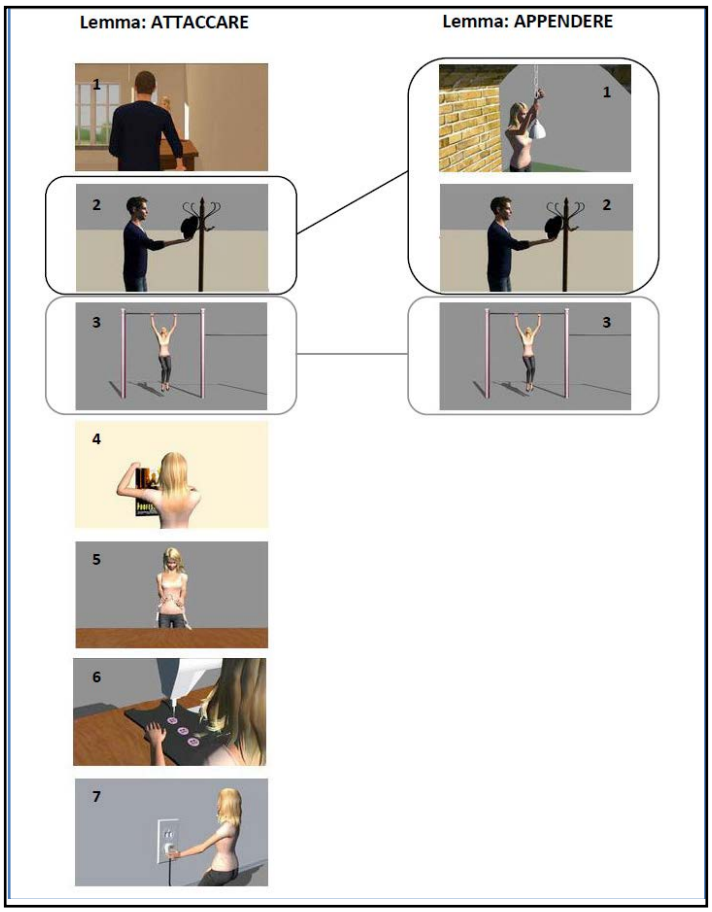

Fig. 3: Classi azionali dei lemmi attaccare e appendere a confronto.

I tipi azionali 1 e 2 del verbo *appendere* sono stati categorizzati come unico tipo (il 2) nel lemma *attaccare*: ciò significa che l'annotatore, tipizzando la variazione primaria del lemma *attaccare*, non ha ritenuto di dover distinguere gli eventi sulla base dello stato risultante del tema (il tema "pende dal riferimento" in 1, e non "pende" in 2); al contrario, il medesimo tratto è stato considerato rilevante per il lemma *appendere*. Eventi che costituiscono più classi azionali all'interno della variazione di un predicato più specifico sono stati insomma categorizzati come unica classe per predicati di maggiore generalità.

3.2 Ipotesi di lavoro

Per gestire casi come quello appena illustrato, nel corso della pianificazione della struttura del database sono state prese in considerazione due ipotesi di lavoro. Una prima soluzione prevede che la granularità dell'annotazione del lemma meno generale venga riprodotta nel lemma più generale. Riprendendo l'esempio in fig. 3, il tratto "sospensione" verrebbe considerato pertinente sia per il lemma *appendere* che per il lemma *attaccare*. Ciò avrebbe come conseguenza il fatto che il database ammetta un'unica tipologia di relazione, l'equivalenza. Ne risulterebbe un DB di notevole semplicità strutturale, in cui è sempre possibile stabilire relazioni 1:1 tra tipi azionali. Un *mapping* così concepito porterebbe però ad una anti-economica sovragenerazione di tipi azionali per i verbi generali.
La seconda soluzione prevede che nella struttura dei dati vengano introdotte relazioni implicite di tipo IS_A. La scelta avrebbe essenzialmente due conseguenze: il database creato conterrebbe gerarchie *implicite* di tipi, e uno stesso tipo potrebbe essere rappresentato nell'ontologia da più scene. A ciò corrisponderebbe un aumento della complessità delle relazioni nel DB; la soluzione, tuttavia, consentirebbe di mantenere contenuto il numero di tipi azionali e soprattutto di garantire l'aderenza della tipizzazione all'intuizione dei parlanti madrelingua.
In ragione della sua maggior coerenza rispetto ai requisiti progettuali, è stata scelta la seconda soluzione.

3.3 Relazioni del DB IMAGACT

Le entità del DB IMAGACT 1.0 sono organizzate mediante due tipologie di relazione:
- Relazione tipo-tipo;
- Relazione tipo-scena.

Nella prima categoria rientra la relazione L_EQ, "*local equivalence*". Nel quadro teorico adottato in IMAGACT (Moneglia 1997) viene definita "equiestensionalità" o "equivalenza locale" la possibilità per due (o più) predicati di applicarsi allo stesso evento o insieme di eventi, sulla base di proprietà di senso. Tale proprietà è rappresentata *indirettamente* nel database dall'appartenenza di una stessa scena alla variazione primaria di due o più lemmi (fig. 4).

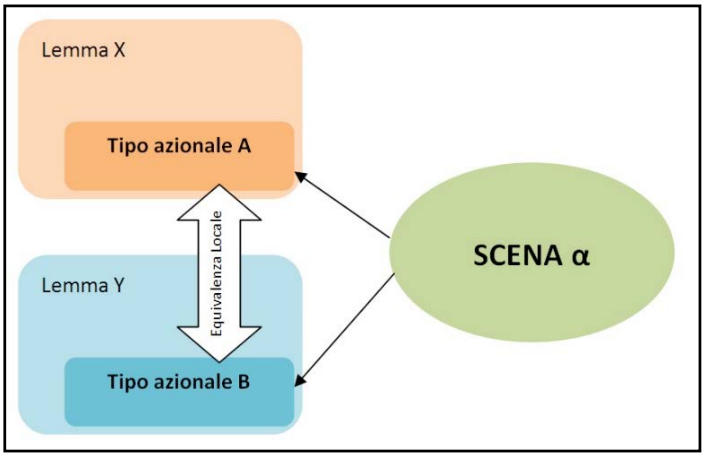

Fig. 4: Relazione di Equivalenza Locale (L_EQ) nel DB IMAGACT.

Dati i tipi azionali a, b, c ed i lemmi X e Y, la relazione ha le seguenti caratteristiche:

- $a \in X, b \in X \Rightarrow \neg L\_EQ\ (a, b)$
- $a \mathfrak{R} b \Rightarrow b \mathfrak{R} a$ (simmetria)
- $a \mathfrak{R} b, b \mathfrak{R} c \Rightarrow a \mathfrak{R} c$ (transitività)

Tipi azionali e scene vengono invece collegati in IMAGACT secondo due modalità (fig. 5):

- PRO, "*prototipo*": la scena è un prototipo per il tipo;
- INST, "*istanza*": la scena rappresenta una possibile realizzazione (non prototipica) della classe di eventi rappresentati in un tipo.

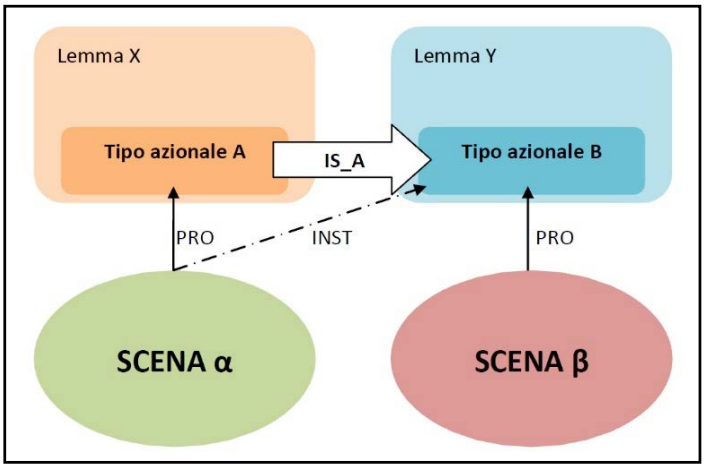

Fig. 5: Relazioni PRO e INST nel DB IMAGACT.

Ogni tipo azionale del database IMAGACT ha associata una, ed una sola, scena con relazione PRO; può invece avere, opzionalmente, più scene connesse con relazione INST.

Ciò corrisponde al fatto che un tipo di azione altamente prototipico per un verbo (es. "*attaccare/appendere* la lampada al soffitto" in relazione alla variazione primaria del lemma *appendere*), possa corrispondere ad una istanza periferica per un altro verbo (l'evento di "*attaccare/appendere* la lampada al soffitto"

è una fra le possibili istanze per il lemma *attaccare*, al pari di "*attaccare*/*appendere* il cappello all'attaccapanni"). Il fenomeno, che non ha natura logica, è connesso alla maggiore o minore marcatezza pragmatica dell'evento e a fattori semantici ancora da investigare.

3.4 Il concetto di "Famiglia di Prototipi"

La scelta di una strategia gerarchizzante ha avuto come effetto l'introduzione in IMAGACT del concetto di "famiglia di prototipi": laddove siano presenti differenze di granularità di annotazione dovute al senso del lemma annotato e tali differenze appaiano consistenti e/o interessanti, classi di azioni distinte all'interno della variazione di un predicato specifico possono essere associate a costituire un unico *cluster* di prototipi in predicati più generali. Con la dicitura "famiglia di prototipi" si intende dunque in IMAGACT l'insieme delle scene connesse ad un predicato allo scopo di esplicitare differenziali linguistici.
In fig. 6 è mostrata la soluzione strutturale adottata per l'esempio discusso nei paragrafi precedenti. Dal punto di vista dell'architettura dell'ontologia, ciascuna scena corrisponde a un nodo della gerarchia.

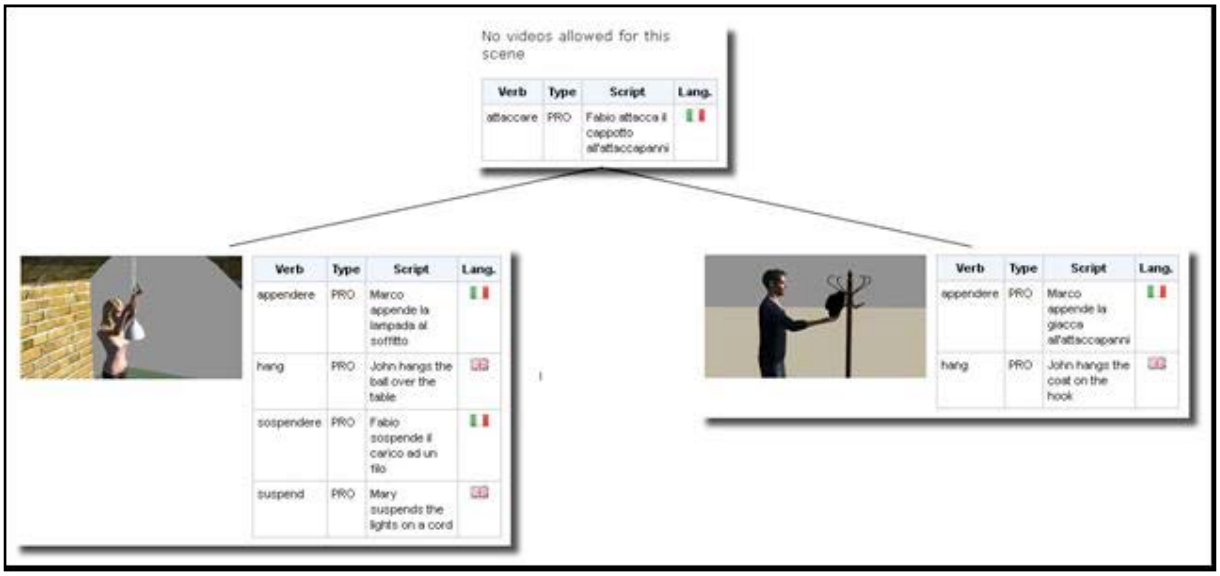

Fig. 6: Esempio di "Famiglia di Prototipi": *attaccare.*

3.5 Troponimi e denominali

Una struttura dati così concepita consente anche di gestire agevolmente varie tipologie di iponimi, e la loro relazione con i tipi azionali dei verbi generali (fig. 7). Tra questi:

- troponimi, ovvero iponimi che esplicitano la modalità con cui l'azione viene compiuta dall'agente (es. *appiccicare* vs. *attaccare*);
- denominali, ovvero iponimi che esplicitano uno specifico materiale o oggetto di cui l'agente si serve per realizzare l'azione (es. *incollare*, *to glue*, *to tape* vs. *attaccare*).

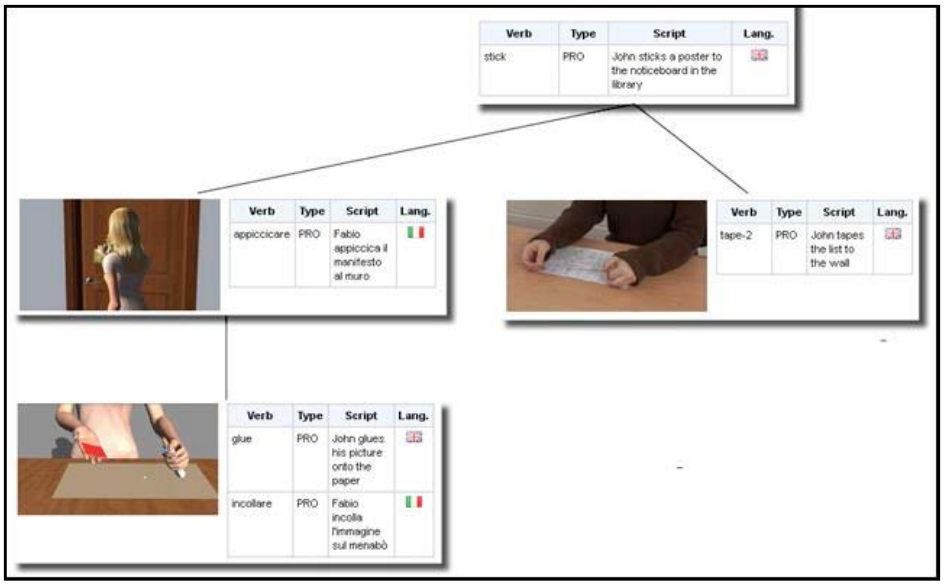

Fig. 7: Esempio di mapping di verbi denominali e troponimi: to stick, appiccicare, to glue, incollare, to tape.

4 Conclusioni

La versione 1.0 del database IMAGACT, rilasciata in data 1/09/2013, contiene 521 verbi ad alta e media frequenza per l'italiano e 550 per l'inglese, connessi ad una galleria di 1010 scene. La risorsa è interrogabile all'URL http://www.imagact.it/.

La metodologia illustrata ha permesso la generazione di tale ontologia al suo stadio attuale: le classi azionali individuate a partire dai lemmi delle due lingue sono state organizzate in una struttura dati coerente, conciliando la necessità di correttezza formale con la volontà di mantenere aderente la tipizzazione della variazione all'intuizione dei linguisti madrelingua che hanno prodotto l'annotazione. Per facilitare l'estensione della struttura dati ad altre lingue, il database è attualmente in fase di revisione e di semplificazione nell'ambito del progetto MODELACT ("From individuation to modelling in natural language ontology of action. Grounding the definition of action concepts on language infrastructures").

5 Bibliografia

Fellbaum, Ch. (1998). *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press.

Kipper-Schuler, K. (2005). VerbNet: A broad-coverage comprehensive verb lexicon. PhD thesis, University of Pennsylvania, Philadelphia, US.

Gagliardi, G. (2014). Validazione dell'Ontologia dell'Azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment (MCI). PhD thesis, Università degli Studi di Firenze, Italia.

Frontini, F., De Felice, I., Khan, F., Russo, I., Monachini, M., Gagliardi, G. & Panunzi, A. (2012). Verb interpretation for basic action types: annotation, ontology induction and creation of prototypical scenes. In: M. Zock & R. Rapp, *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, CogALex III*, The COLING 2012 Organizing Committee, pp. 69-80.

Moneglia, M. (1997). Prototypical vs. not-prototypical verbal predicates: ways of understanding and the semantic types of lexical meanings. In: *Vestnik Moskovkogo Universitatea (Moscow State University Bulletin)*, 2, pp.157–173.

Moneglia, M. (*in press*). The Semantic variation of action verbs in multilingual Spontaneous speech Corpora. In: T. Raso, H. Mello (eds.), Spoken Corpora and Linguistics Studies, Amsterdam: Benjamin.

Moneglia, M. & Panunzi, A. (2010). I verbi generali nei corpora di parlato. Un progetto di annotazione semantica cross-linguistica. In: I. Korzen & E. Cresti (eds.) *Language, Cognition and Identity. Extension of the Endocentric/Esocentric Typology*. Firenze: FUP, pp. 27-46.

Moneglia, M., Monachini, M., Calbrese, O., Panunzi, A., Frontini, F., Gagliardi, G. & Russo, I. (2012). The IMAGACT cross-linguistic ontology of action. A new infrastructure for natural language disambiguation. In: N. Calzolari et al. (eds.), *Proceedings of the Eigth International Conference on Language Resources and Evaluation – LREC'12*, pp. 948-955.

Rosch E. (1978). Principles of Categorization. In: Rosch E. and Lloyd B.B. (eds.), *Cognition and Categorization*. Hillsdale, NW: Erlbaum. 27-48.

Acknowledgements

Il progetto IMAGACT è stato finanziato dalla regione Toscana nell'ambito del programma PAR.FAS. (linea di azione 1.1.a.3) . Ulteriori ricerche sul database IMAGACT, incluso questo articolo, sono state realizzate grazie al contributo del progetto MODELACT (2013-2016), finanziato nell'ambito del programma nazionale Futuro in Ricerca.