Bootstrapping the semantic variation of general verbs from spontaneous speech corpora. The IMAGACT ontology infrastructure

Alessandro PANUNZI (Università di Firenze)

In all language modalities Action verbs bear the basic information that should be processed in order to make sense of a sentence. Especially in speech, they are the most frequent structuring elements of the discourse, but their semantic nature does not specify the referred action making puzling natural language undestanding tasks. In spoken languages the most frequent action verbs are "general", i.e. they are able to extend to actions belonging to different ontological types.

The semantic variation of verbs in the language use is well known. For instance in the more influential lexical data base, Wordnet, the verb "to open" records 11 synsets

WordNet- open (verb)
1- S: (v) open, open up (cause to open or to become open) "Mary opened the door"
2- S: (v) open, open up (start to operate or function or cause to start operating or functioning) "open a business"
3- S: (v) open, open up (become open) "The door opened"
4- S: (v) open (begin or set in action, of meetings, speeches, recitals, etc.) "He opened the meeting with a long speech"
5- S: (v) unfold, spread, spread out, open (spread out or open from a closed or folded state) "open the map";
6- S: (v) open, open up (make available) "This opens up new possibilities"
7- S: (v) open, open up (become available) "an opportunity opened up"
8- S: (v) open (have an opening or passage or oet) "The bedrooms open into the hall"
9- S: (v) open (make the opening move) "Kasparov opened with a standard opening"
10- S: (v) afford, open, give (afford access to) "the door opens to the patio"; "
11- S: (v) open (display the contents of a file or start an application as on a computer)

However the crucial semantic variation is not clearly identified. The instruction "open x" can lead qualitatively different actions that are not listed in the above repository. For instance a)"opening a window", b)"opening a box", c)"opening the umbrella". The action "opening a box" characterizes by the access to some inside space, while there is no inside space in the action "opening a window", but rather an outside space. There is neither inner nor outdoor space in the case of "opening the umbrella". In other words from a cognitive point of view the model of the referred action changes radically in those cases and does not correspond to one single action type, but rather to various distinct types.

The onset of these action types within the variation of the action verb "to open" is not a consequence of phraseological or metaphorical usages of the verb, as in most of Wordnet synsets. Action types a), b) and c) instantiates the verb "in its own meaning" and no type is more in the core extension of the verb than the others (Rosh, 1978). In particular each type is productive; i.e it varies over instances of the type: a) opening a window, a door, the tent. etc; b) opening a box, a suitcase, the tea-pot; c) opening the umbrella, the pen, the map. We call "general" those action-oriented predicates that can refer, in their own meaning, to many different action models (see. Wittgenstein 1953; Givon 1986; and Lakoff, 1987 for the early identification of this phenomenon in natural languages).

General predicates are the most frequent verb class in speech, both in terms of occurrences and in term of lemmas, and correspond to the linguistic categorization of the most frequent actions of everyday life (Moneglia & Panunzi 2007). Moreover, each language categorizes

action in its own way and the cross-linguistic reference to everyday activities is also unpredictable. For instance in some Romance Language breaking events like "braking a nut, an egg, etc" are possible extension of the predicates corresponding to to open, but this type is excluded in English (Bowerman, 2005). The variability of crosslinguistic reference to action cause major problems for human language processing and translation tasks.

If the application of general verbs to the action types in their extension is productive than the linguistic categorization of a type should be in principle predictable. But the ontology of action is not available in any existing repository and the actual variation of general verbs even in the more common languages is unknown. Despite the richness of the information provided by current ontologies, there is no guarantee that the actual semantic variation of action verbs in the ordinary use of language is recorded and moreover the "productive variation" is not split from metaphor and phraseology.

Spontaneous Speech Corpora, which are now available for a lot of languages, contain both the reference to the more frequent actions in everyday life and their lexical encoding and can be used to bootstrap this information.

The paper will show the incidence of general verbs in the high frequency verbal lexicon of spontaneous speech corpora (BNC for English and C-ORAL-ROM for the Romance Languages) and will present the IMAGACT annotation infrastructure aimed at bootstrapping the semantic variation of general verbs from BNC Spoken and C–ORAL-ROM through annotation procedures. IMAGACT will use corpus-based and competence-based methodologies for simultaneously extract from such corpora both the referred action types and their linguistic encoding in the various represented languages.

The corpus-based strategy relies on a bootstrapping procedure which split the metaphorical and phraseological usages from proper occurrences and then classify proper occurrences into types.

The key innovation of IMAGACT is to provide a methodology which exploits the language independent capacity to appreciate similarities among scenes, distinguishing the "Identification of action types" from their "Definition". Only the identification is required to set up cross-linguistic relations. In Wittgenstein's terms, how can you explain to somebody what a play is ? Just point out a play and say "this and similar things are plays" (Wittgenstein, 1953).

To this end the ontology building makes use of the universal language of images which allows reconciling in an unique ontology the descriptions derived from the annotation of corpora belonging to different languages.

The project will result in an Inter-linguistic Action Ontology derived from corpus annotation, and will allow the mapping of action types onto the corresponding predicates in the implemented languages.

References

BNC [online] http://www.natcorp.ox.ac.uk/

Bowerman, M. 2005. Why can't you "open a nut" or "brake a cooked noodle". Learning cover object categories in Action word meanings. In L. Gershkoff-Stowe, D. H. Rakison, Building Object Categories In Developmental Time. New Jersey: Lawrence Erlbaum Associates.

Cresti E, Moneglia M. (eds) 2005. C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages. Amsterdam: Benjamins

Givon, T. 1986. Prototypes: Between Plato and Wittgenstein. In C. Craig (ed.) Noun Classes and Categorization. Amsterdam: Beniamins. 77-102.

IMAGACT http:lablita.dit.unifi.it/projects/imagact

Lakoff, G. 1987. Women, Fire, and Dangerous Things. What Categories Reveal about the Mind. Chicago/London: University of Chicago Press.

Moneglia, M., 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, Proceeding of the International Workshop on the Semantic Representation of Spoken Language (SRSL7). Salamanca: Universidad de Salamanca.

Rosch E. 1978. Principles of categorization. In E. Rosch, B.B Lloyd (eds) Cognition and categorization. Hillsdale (NJ): LEA. 27-48.

Wittgenstein, L. 1953. Philosophical Investigations. Oxford: Blackwell.

WordNet http://wordnet.princeton.edu/