Action Predicates and the Ontology of Action across Spoken Language Corpora. The Basic Issue of the SEMACT Project

No Author Given

No Institute Given

1 General and activity verbs

An action is a pattern of world's modifications by an actor that can be applied to an open set of objects. Actions and objects are ontological entities that necessarily co-occur in the world, but each action is independently defined from the set of objects to which it is applied. In short each action can be applied to an open set of objects and each object can undergo an open set of actions.

This means that actions are productive in the world. Classical works in Psychology and Philosophy, but also in Computers science and Artificial intelligence, roughly agree on this conception (Von Wright, 1963; Minsky, 1968; Tomasello, 2003). The productivity of action is mirrored at the linguistic level. Assuming selection restrictions, a predicate referring to an action can be applied to an open set of arguments and such arguments also undergo an open set of action predicates. Indeed, the ability to use natural language predication emerges in the child specifically when both conditions are satisfied.

However natural languages predicates go beyond this basic productivity of actions. For instance, considering the pragmatic circumstances of the world, the act which corresponds to the instruction "to open" can lead to qualitatively different events: (a) "opening a window", (b) "opening a nut", (c) "opening the umbrella".

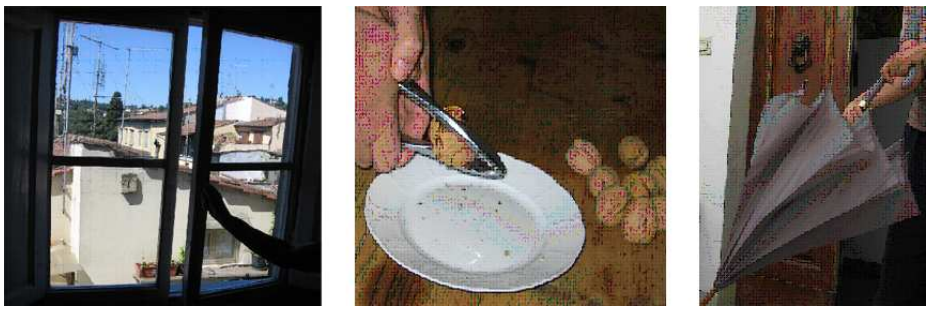

Fig. 1. (a) opening a window, (b) opening a nut, (c) opening the umbrella.

According to the judgment given by humans on the basis of their cognitive capacities, in the above circumstances more then one single action type occurs. Indeed it is surprising that so different events fall within the extension of the same predicate.

This paper will briefly sketch this intriguing property, that regards the main part of the action oriented verbal lexicon of natural languages, and will focus specifically on its consequences for linguistic translation. The relation between action types and action predicates will then be considered from the perspective of spontaneous speech corpora. We will argue that the semantic annotation of multilingual corpora of spontaneous speech is the proper method for the induction of such data and that this task, that is the main concern of the SEMACT project, is feasible.

The judgment that Fig. 1(a-c) corresponds to three different event types is strong and comes from those evidences that are embodied in the pragmatic of action that are significant to human cognition. Despite the fact that the form of events is radically underdetermined, we can try to figure out, at least informally, the reason for this judgment. The action of "opening a nut" allows the access to some inside content, that is not directly available, while there is no inside content in the action "opening a window", but rather an outside space. In the case of "opening the umbrella", there is neither content nor out-space.

In other words, the action performed in Fig. 1(a-c) changes in crucial properties and for this reason three distinct types rather that to one single action type arise. This is confirmed by the productivity of each action type in accordance with the basic properties of actions models. For example a human is also able to judge that, on the contrary, the same type of action is performed when the predicate "to open" is applied to the objects listed in the sentence: *Bill opened the door / the fencing / the tent*, and the resulting actions belong to the same type represented in Fig. 1(a).

The same occur with the set of objects in the sentence: *Bill opened the egg / the package / the TV-case*, that give rise to events belonging to the type of Fig. 1(b) and also with the objects in the sentence: *Bill opened the deckchair / the biro / the lock*, that instantiate the events type in Fig. 1(c).

The similarity among instances in each of the three series can be appreciated regardless the variation of the properties involved. For example no one would imagine that the word "to open" refer to different action models when comparing "opening the door" and "opening the fencing". Therefore is safe to say that an Action-predicate like"to open" can be applied to actions which belong to different types. In other words the relation between Actions and Action-verb is not a one to one correspondence and, in this domain, ordinary language, it does not mirror the ontology of action.

Although natural Language Semantics is not frequently concerned with it, the variation of action predicate in natural language should be well known. The argument regarding the pragmatic variation of the events corresponding to the Deverbal Noun "Play" is the main historical antecedent of the evidence just presented. More specifically, Wittengstein (1953) used the properties of the rad-

ical variation of the predicate "Play" to demonstrate that natural concepts are not strictly governed by semantic rules and introduced the notions of prototype and family-resemblance to explain, in principle, how natural predicates can be extended (Givon, 1986).

In this paper we will not be concerned with the semantic explanation of this linguistic property but, nevertheless, we can observe that "to open" applies in its own meaning to the models in Fig. 1(a-c). Moreover in none of such instances "to open" is used in a more appropriate way than in the others. In other word there is no action in the above set that can be considered "prototypical".

However, if categorization phenomena are seriously considered, we can notice that this phenomenon does not regard all action predicates of a given natural language. Activity verbs like "to eat" or "to run" or "to iron" (Dowty, 1979) strictly refer to just one action type:

Fig. 2. (a) eating the pork chop, (b) eating the ice cream, (c) eating the soup.

Indeed also the properties embodied in the above instances of "to eat" vary consistently. For example in Fig. 2(a) teeth are strongly involved and the actor must chew with care. In Fig. 2(b) the actor performs the action with tongue. In Fig. 2(c) neither teeth nor tongue are involved. However, contrary to what we have seen in Fig. 1(a-c), humans do not show any interest to consider the events in Fig. 2(a-c) as a distinct event types. Despite the variation of important properties, the latter events are categorized in the same manner; i.e. always the same action is performed.

In short, in this case linguistic categorization and conceptual categorization go hand in hand and a nice one to one correspondence between linguistic predicates and actions types occurs. Considering the conceptual categorization of the various instances of action oriented predicates we call "general verbs" those action-oriented predicates that refer to many different action types, while we use the traditional term "activity verb" to indicate those action oriented verbs that do not show this property.

2 Cross-linguistic correlations of action-verbs

Although all languages have a general predicate lexicon, the ratio between the two classes may change. Most languages seem like to categorize action more frequently with general verbs, while other (e.g. Corean, Danish and probably German) have the opposite tendency (Choi & Bowerman, 1991; Korzen, 2005).

General predicates constitute probably an advantage for linguistic categorization at the intra-linguistic level, but this linguistic device makes for sure more problematic the inter-linguistic reference to actions, even for what regards common activities of everyday life. More specifically, general predicates cause the following phenomena in establishing correspondences in bilingual dictionaries: (a) given two languages, there is no necessary correspondence between the action types that are in the extension of verbs of those languages and therefore, no one-to-one translation between general predicates in different languages can be established; (b) activity verbs in different languages are almost in direct translation relation.

Given two languages, once an action type is categorized through an activity verb in both languages, then the corresponding predicates find a direct translation; i.e. they are applied in parallel to the same set of pragmatic variations. For example the variation in Fig. 2(a-c) can be referred by the English verb "to eat" and, in parallel, by the Italian verb "mangiare", the French verb "manger" and so on. This is frequently not the case with general predicates. For example both English and Italian use one high frequency general verb to identify the action type instantiated in Fig. 3(a), i.e. the entries "to take" and "prendere", that, in accordance with bilingual dictionaries are in a translation relation.

However the set of action types that are in the extension of these two predicates are not in one to one correspondence. English apply "to take" also to the action type in Fig. 3(b), while Italian is forced to describe this event with the general predicate "portare". However the semantic competence of English does not allow the application of "to take" to most action types referred by the Italian predicate "portare". For example in Fig. 3(c), that is one of the central instances of the Italian verb "portare", only "to bring" will find application. On the other side the Italian verb "prendere" is also extended to the actions types in Fig. 3(d), where on the contrary English apply a second general predicate "to pick up".

Therefore there is no one to one translation relation between general action predicates in different languages, and what is more relevant, from a pragmatic point of view we observe the intersection of the extension sets of many general predicates. The semantic partition between general and activity action predicates in natural language lexicon have a strong impact at cross-linguistic level and brings to puzzling translation problems. These problems do not belong to the realm of phraseology, but rather to the productive aspect of the semantic competence and their solution crucially involves the identification of the pragmatic variation of action oriented predicates. But the pragmatic variation of general verbs in different languages is at present unknown.

Fig. 3. (a) to take/prendere, (b) to take/portare, (c) to bring/portare, (d) to pick up/prendere.

The peculiar nature of the linguistic categorization of action has a strong impact on multilingual translation even for quantitative reasons. As will see below, general predicates are the more frequent semantic category in the verbal lexicon and therefore they cause a high number of unpredictable translation contexts. The follwing section will sketch the frame of a corpus based cross-linguistic research (SEMACT) that figure out the induction of this variation from spontaneous speech data.

3 Action-verbs and spoken language corpora

The linguistic reference to actions and the relevance of those actions in our life go hand in hand. The actual use of Action oriented verbs in the linguistic performance can therefore be appreciated observing their occurrence in spontaneous speech corpora. These corpora, now available for the main languages of the EU, document how common actions are represented in the various languages and therefore they constitute the main source of information for what regard the semantic variability of the lexicon.

The induction of data regarding the variability of action verbs in different languages requires a clear research strategy for both practical and theorethical reasons. The verbal lexicon occurring in large corpora records over 5,000 items per language and the study of their pragmatic variability, that requires competence based annotation, is not feasible. Moreover this lexicon is not homogeneous and the selection of the relevant domain of investigation is needed.

The first strategy adopted in SEMACT is to concentrate on those entries that are the more frequent in each corpus and that are therefore responsible for the language specific categorization of highly probable events of the everyday life. A second strategy is to select, within the high frequency lexicon, those verbs that strictly refer to actions that constitute the primary universe of human beings: i.e actions whose linguistic categorization determines the onset of a language specific semantic competence in early acquisition (Tomasello, 2003). We will see that this lexicon also constitute the most part of the language use in spontaneous speech corpora.

According to pilot studies on spoken Italian corpora , general predicates are the most frequent verb class in spontaneous speech, both in terms of tokens and

in terms of lemmas. This result has been replicated on larger corpora of various languages of the EU in the preparatory studies of the SEMACT project[1]. The data presented below identify the number of action oriented verbs belonging to the high frequency lexicon of Italian (1284978 total tokens) and English (10378225 total tokens) corpora and show the incidence of this lexicon in the speech performance at cross-linguistics level.

In order to identify a significant portion of the high frequency verbal lexicon in the two language corpora the following criteria have been applied: (a) the verbal entries falling within the fundamental lexicon, that is the set of lemmas that cover 90% of tokens in each corpus, have been identified in each resources; (b) this lexicon have been enriched with an additional series of highly ranked verbs immediately following the fundamental verbs in the frequency lists (in the proportion of roughly 1/3 of the original set).

The Italian and English resources record respectively 17,646 and 40,583 recognized lemmas. Despite this huge discrepancy due to the bigger size of the English corpus, the fundamental lexicon of both resources turns out, very nicely, strictly comparable. In the English resource the fundamental lexicon records 1588 entries, containing 287 verbs. In the Italian resource it records 1590 entries, containing 299 verbs[2]. Adding the additional series, two equal sets of 405 verbs have been derived from the frequency lists of the two corpora. These verbs can therefore be considered, given the resources available nowadays, the most frequent verbal lexicon characterizing English and Italian in their spontaneous speech use.

High frequency verbal lexicon can be classified both in terms of number of lemmas and in terms of percentage of occurrence in the corpora[3]. Action oriented lexicon is first compared to the verbal lexicon that does not instantiate actions. Subordinatives, by definition, do not represent actions. Data from the Italian and English corpora show a strict comparability. The number of lemmas in the fundamental lexicon that identify actions is higher with respect to subordinative verbs (5/3 relative factor). The occurrences of Verbs in the corpus are almost equally divided into Action oriented verbs and Subordinative verbs (dicendi, sentiendi, putandi, performatives, circumstantial ect.). That means that, in spontaneous speech contexts, 50% of the occasions in which a verb is used, an eventuality is also represented.

Action oriented verbs have been also divided into classes, in order to produce a typology that is significant, from both a cognitive and a linguistic perspective, for the study of the language specific categorization of action. To this end the

[1] The spoken section of the BNC is the main available source for English. For what regard Romance Languages, the C-ORAL-ROM corpus constitutes the main freely available source for French and Portuguese. For Italian and Spanish larger collections of spoken resources are available and have been used for this research (C-ORAL-ROM, LABLITA, CORLEC and LIP).

[2] Similar values have been recorded for French, Spanish and Portuguese.

[3] The occurrence of modals, auxiliaries, copula, and other verbs whose function is mainly structural (like *to have* and *to do*), have not been considered in the statistics (roughly 50% of the total verbal use in spontaneous speech).

research concentrate on those actions, which, like the previous ones, constitute the very basic frame of the interaction between humans and the world. Therefore various activities and eventualities have not been considered. More specifically the following types have been gathered in a spurious set (OTHZ): Social activities (e.g.. to buy, to rent, to welcome); activities that refer to the Dialogical exchange (e.g. to quote, to discuss); Abstract actions (e.g. to modify, to derive, to enhance); States (e.g. to belong; to cope, to accord).

Finally the resulting verb classes to be considered identify respectively General verbs (GEN) , Activity verbs (ACT) and Movement verbs (MOV).

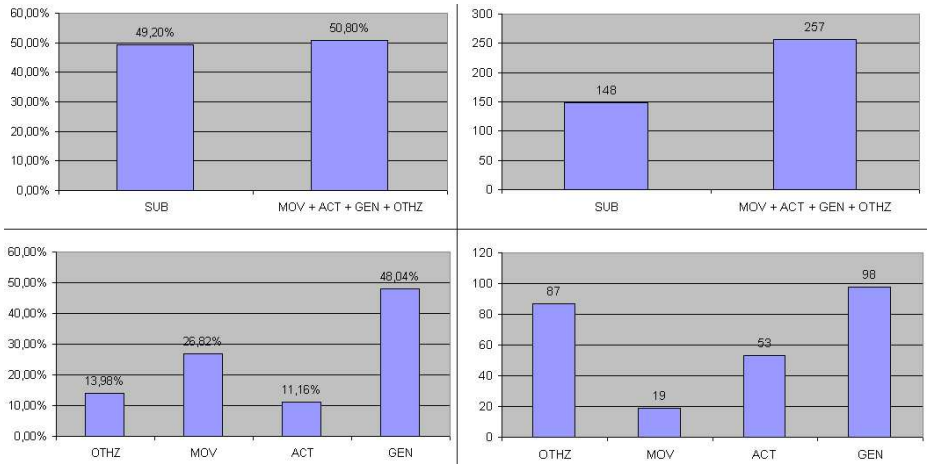

Fig. 4. % of tokens and n. of lemmas in high frequency verbal lexicon (English).

Within the set of Action oriented verbs MOV, ACT and GEN, which identify basic actions, record together the great majority of tokens. Moreover histograms show that this subset, that we assume to be responsible for the linguistic categorization of everyday basic actions is constituted by a restricted number of lemmas (170 for English, 158 for Italian). Looking to the action verb classes, English and Italian, are both characterized by a verbal lexicon oriented to "general predicates". This class records two times the entries of "activity predicates", and five times the entries of "movement verbs". The two languages are therefore similar for what regard the structure of the fundamental lexicon in this respect.

The relative probability to refer an action through the linguistic categorization conveyed by a general verb is high: the use of general verbs is from 3 to 5 time higher than the use of activity verbs and from 1,5 to 2 time higher with respect to movement. The number of entries referring to basic actions of the everyday life and their incidence on the speech performance is a crucial datum. Corpus based analysis show that this number is restricted at cross-linguistic level and that their incidence is high. A corpus based research of their variation is therefore a possible objective.

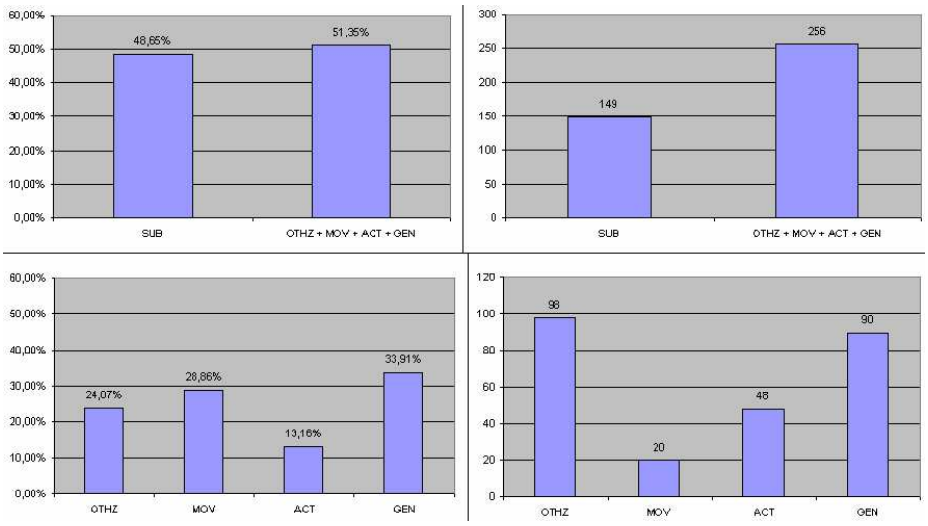

Fig. 5. % of tokens and n. of lemmas in high frequency verbal lexicon (Italian).

References

BNC. http://www.natcorp.ox.ac.uk/

Choi, S., Bowerman, M. 1991. Learning to express motion events in English and Korean: the influence of language specific lexicalization patterns. Cognition, **41** 83–121.

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

De Mauro, T., Mancini, F., Vedovelli, M., Voghera, M. 1993. LIP. Lessico di frequenza dellitaliano parlato. Milano: ETAS.

Dowty, D. 1979. Word meaning and Montague grammar. Dordrecht: Reidel.

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

Korzen, I. 2005. Endocentric and esocentric languages in translation. Perspectives. Studies in translatology, **13(1)** 21–37.

LABLITA http://lablita.dit.unifi.it/corpora/

Marcos Marn, F. 1992. CORLEC. "El Corpus Oral de Referencia de la Lengua Espanola Contempornea". Informe del proyecto. Madrid. Accesible a travs de ftp://ftp.lllf.uam.es/pub/corpus/oral.

Minsky, M. 1968. Semantic Information Processing. Cambridge (MA): MIT Press.

Tomasello, M. 2003. Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge (MA): Harvard Univerisity Press.

TreeTagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

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

Von Wright, G. H. 1963. Norm and Action: A Logical Enquiry. London: Routledge & Keegan.