Words on the edge: conceptual rules and usage variability

Elisabetta Gola* e Stefano Federici*

1. Introduction

The high degree of variability in language use, especially in non literal utterances, represents a problem for formal semantics and for the attempts of simulation of the linguistic behavior in the field of Artificial Intelligence. Natural Language Processing (NLP) systems, in fact, have been until now heavily based on compositional mechanisms, that is mechanisms that assign meaning to complex constructions based both on the meaning of their components (meanings that are listed in a lexicon) and on restrictions on how the meaning of these components interact. Because of this choice -both conceptual and by design- metaphors and other forms of non literal meanings have been considered an inescapable problem and an obstacle by computational systems based on those principles (Carbonell 1982).

In this paper, we describe a corpus-based method that gets along without deep compositional analysis and try to cope with the problem of variability in metaphors and the way in which they systematically appear in ordinary language and frames our thought.

2. Background

Metaphor has been studied and analyzed from several perspectives that are so different that can be considered almost opposite. This (apparent or real) contrast, that characterized the analysis of figurative phenomena, has highlighted the role of metaphors either as a phenomenon of language (that is a surface communicative structure, Genette 1972) or as conceptual structures (cognitive semantics, Lakoff & Johnson 1980). This dichotomy generated further dichotomies such as the dichotomy between conventional metaphors (*dead* metaphors belonging to the lexicon and settled in language) and creative metaphors (*live* metaphors created in original visions with respect to established knowledge).

These dichotomies led to further issues: one issue is the fact that compositional rules are systematically violated in metaphors. Just to give an example, in a compositional system it is difficult to derive the right meaning of the Italian sentence "*guidare un'azienda*" (to lead a company) if the system doesn't explicitly list in the lexicon the knowledge that even a company can be "*guidata*" (led, literally "driven"). The other issue is related to some difficulties of the theory of conceptual metaphor in predicting how metaphors are effectively used in texts (Hardie *et al*. 2007).

To try to solve this complex problem, computational linguists have been working on the role of metaphors in Natural Language Processing systems, highlighting the need to face at least three aspects: lexical issues (description of lexical structure needs to be integrated with non literal properties (Moravcsik 2001)), world knowledge representation (ontologies have to include metaphorical mapping among different areas (Martin 1990)) and effective use in texts (particular -but often frequent- cases can only be found in real texts (Gola 2005)). Nevertheless, we still do not have a comprehensive theory that could represent the basis for NLP models.

In our opinion, a comprehensive theory of how metaphors can be integrated in NLP systems must follow a corpus-based approach and it should be able to analyze at the same time generalization and variation of metaphorical uses. The rationale behind this is the consideration that, when looking at metaphors from a conceptual point of view, probably

* University of Cagliari.

many generalizations are found, but the variation and variability of non literal expressions are often ignored.

3. Corpus-based approaches to metaphors

Many studies have shown the limits of conceptual approaches to the analysis of metaphors, especially because the adopted frameworks lack a satisfying empirical analysis (see, for example, Deignan 2005). As pointed out by Keller et *al. (2008)* «rather than restricting the empirical application of conceptual metaphor theory to single data sets, corpus-based metaphor analysis seeks to ascertain metaphor usage in larger text collections that represent the various voices of a discourse community» (p. 142).

This means that «[t]o date, most of the works have taken a lexical approach of some description and analyzed concordances of pre-defined search strings of single words or members of word fields» (Keller *et al.* 2008, p. 143), combining some manual and computer-assisted analysis (see for example the study devoted to plant metaphors by Deignan 2005: pp. 174-183). The analysis in these cases, being limited to predetermined search strings, can't handle new cases.

The machine learning approach that we embrace in this paper also relies on some manual input and analysis in an attempt to overcome the boundaries between different level of analysis so to evolve and grow along corpora change.

4. Contest of the problem

Our proposal combines the corpus-based approach with an analogy-based machine learning engine that can spot patterns of similarity among texts that have only been tagged in a general (i.e. non goal-specific) way (Fig. 1).

To give just an example, in our system, if we want to express the knowledge that a given sequence of syntactic categories is correct, we don't need to state a specific rule, but it is sufficient to tag a small number of occurrences with respect to their specific morphosyntactic properties and then run the analogy-based engine. The engine, basing on the combinations of properties it spotted in the learned texts, decides which unknown combinations of properties are acceptable and which are not. This will be of invaluable help for the next step of semantic analysis.

When working on the analysis of metaphor variability, we propose to analyze them by combining a phase of semantic tagging (that will add a light structure to the corpus by means of domain information) to the position of the words in a semantic net, organized in an ontology that includes information on metaphorical conceptual mappings (Lakoff and Johnson 1980, 1999; Martin 1980) stating, for example, that between the domain of animated things and the domain of ideas it does exist a high probability of metaphorical mapping. So, during the corpus-based analysis phase, the analogy-based engine will be able to organize the results by identifying similarities on the basis of contexts containing the kind of information described above.

For example, if the system has learned that the verb "to die" is associated to animated subjects and that animated subjects metaphorically project their properties on ideas, a sentence like "*La loro idea è morta sul nascere*" (literally, "their idea was dead since the beginning") is considered as perfectly acceptable (contrary to what a system based on formal semantics would do). This kind of analysis would allow to be applied to all variations of a conceptual metaphor and would also allow to understand its extension, its degree of conventionality, its level of "purity" with respect to mixed metaphors.

This kind of analysis would have a great impact on the validation of cognitive semantics theories of conceptual metaphors. Indeed, these theories have been usually formulated more on the basis of a limited set of examples and on speaker's intuition than on the analysis of broad collections of real texts. This is due to the lack of a powerful and flexible computation tools that would fit this specific goals.

Indeed, when reading a text sequentially, metaphors are found that appear apparently different, with a very high degree of variation. Instead, following a corpus-based approach combined with a machine learning engine, it is possible on the one hand to determine the degree of conventionality without having to decide it case by case; on the other hand different degrees of variation can be highlighted with respect to the underlying similarity.

5. Methodology

This path has been proposed in an alternative approach to NLP based on a machine learning approach (Federici 1991). The theoretical framework of this perspective refers to computational systems that are based on the ability to learn rules by themselves, as in autopoietic systems (Livingston 2006).

The base for the learning process is a *boot-strapping* mechanism in which an initial phase of manual intervention on a small part of the corpus is followed by a phase of automatic learning and extension, to unknown data, of the knowledge acquired in the learning phase. A boot-strapping process is indeed characterized by the iterative application of automatic tools starting from a small set of manually analysed data. At each cycle, manual analyses are automatically assigned to raw data and manually revised (fully or even by sample), so to have a good starting point for the next automatic cycle.

Then the boot-strapping cycle is reapplied by allowing the AI algorithm to assign one of the possible interpretations to a further set of the remaining contexts (*extension*). The automatically assigned interpretations are then subsequently manually revised and corrected (if necessary). And the process is started over and over until all occurrences in the corpus are automatically assigned and manually revised.

In this paper we will show that, in this view, metaphoric principles are naturally part of the computational learning process. By showing this we will also see that literal meanings are not to be considered privileged in some way with respect to non literal ones (Gola & Federici 2000).

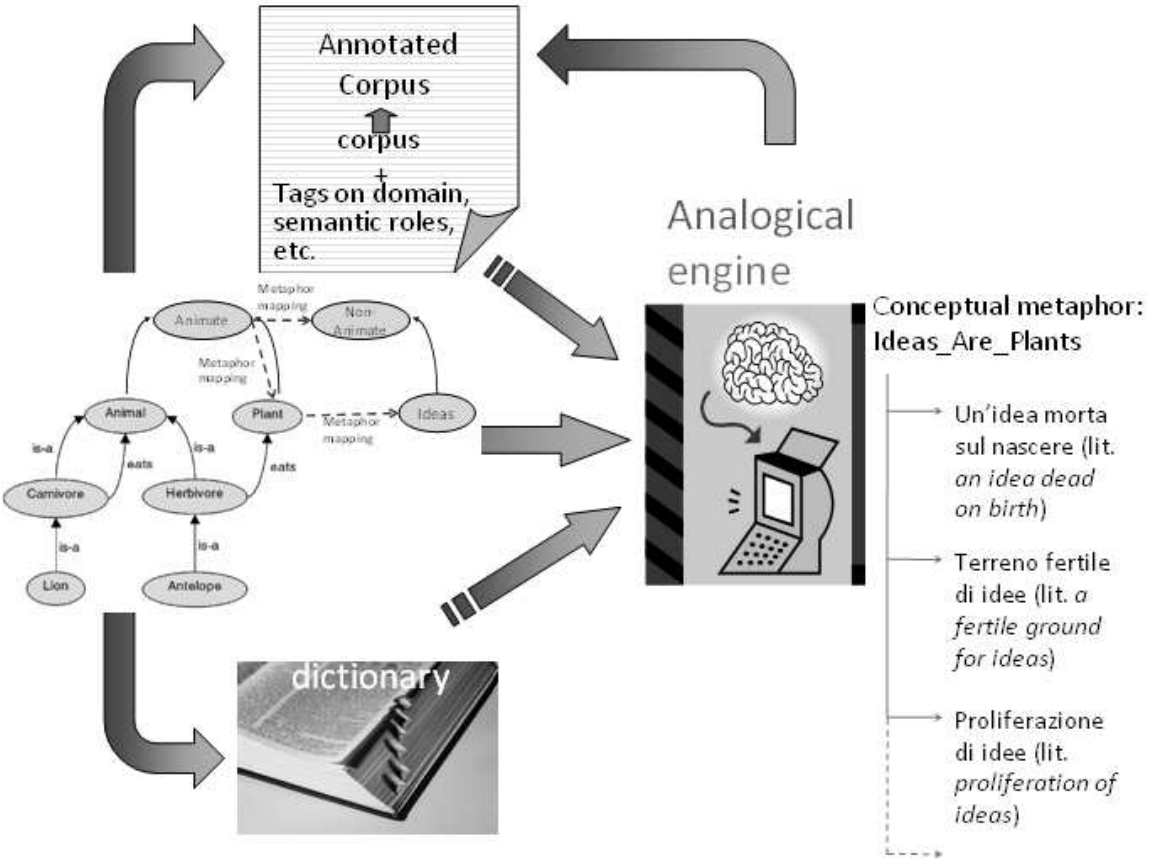

Figure 1. A representation of the process

6. Case study: Analysis of the verb "guidare"

Experimental results supporting this hypothesis are reported from a study based on approximately 800 occurrences of the verb "guidare" (to drive) and other verbs related to the verb "guidare", that is verbs which are used in combination with nouns that co-occur with "guidare". These co-occurrences have been extracted from the Garzanti Italian Dictionary, while the corpus has been collected from an 80 million chars subset of the Italian online newspaper "La Repubblica" (2001-2005).

It can be shown that by acquiring linguistic knowledge from real textual data, metaphorical occurrences (that often represent more than 70% of the total amount of occurrences, Gola 2008) are used to extend their meaning to other, unknown metaphorical usages (Fig. 2).

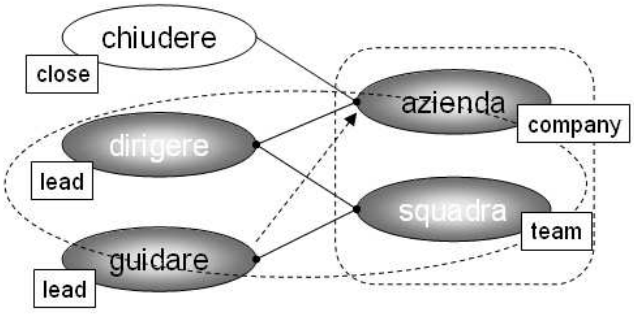

Figure 2. An example of the analogical extension.

Literally, the verb "guidare" means *drive a vehicle*: a car, a truck, a train, a motorcycle, etc. According to compositional perspectives, to be able to semantically understand this predicate we should represent it through some lexical information, restrictions, and rules. As for the structure, the verb 'guidare' is a predicate with two roles: a subject of the predicate (the driver) and an object (a vehicle). Some additional roles can be a location, the type of movement, the time, etc. For the sake of simplicity, we will refer to mandatory complements/roles only. If we adopt semantic restrictions or selectional preferences to guarantee the correctness of well-formed sentences, we should state that the object of "guidare" must be of type "vehicle". All other cases would be then considered as anomalous ones. So, a sentence like "guidare l'auto" (*drive the car*) would be considered as correct, while "guidare una fragola" (*drive a strawberry*), "guidare una lampadina" (*drive a light*), etc. would be discarded as incorrect.

But the real world is not like that and it is not easy to split meanings in separate classes. Let's see the following sentences:

(1) "un percorso che guida alla scoperta dell'opera di questo artista straordinario"
(*a path that guide to the discovery of the works of this extraordinary artist*)
(2) "Quali apparecchi crede che guideranno la diffusione dei giornali online? I Pc."
(*which instruments do you think will drive the spreading of online journals? PCs*)
(3) "Tre obiettivi ne guidano la strategia."
(*Three objectives drive its strategy*)
(4) "Così Trapattoni guida la nobile avanguardia del calcio italiano"
(*In this way Trapattoni lead the noble forefront of Italian soccer*)

Looking at these 4 sentences, it is manifest the difficulty of the compositional perspective outlined above: in three sentences out of four the subject is not a human being (path, instrument, objective) and in the four sentences above the object is not a vehicle. But, we are perfectly able to understand those sentences, exactly as we understand:

(5) "A Grosseto Schumacher guiderà la Ferrari".
(*In Grosseto Schumacher will drive a Ferrari*)

Still, we could think that sentences like (5) can play the role of semantic benchmark. But, in which sense? If we look up the description of the meaning of "guidare" in whatever a dictionary, we immediately realize that the main senses of this verb, even if is related to physical and material activities, are actually metaphorical.

Indeed, the meanings of 'guidare' can be brought back to four core meanings[2], three of which are not literal:

SIGNIFICATO	***CARATTERISTICHE***	***ESEMPIO***
FAR DA GUIDA (synonyms = condurre (*lead*), dirigere (*run*), ammaestrare (*train*))	Non material metaphor	"Suo padre lo aveva sempre guidato tra le intemperie della sua vita" (*his father always guided him through the troubles of life*)
COMANDARE (synonyms = reggere (*govern*))	Lead metaphor	"Il generale guidò con coraggio l'esercito alla conquista" (*The general led his troops to the victory*)
INDIRIZZARE (synonyms = istradare (*guide*), regolare (*settle*))	Material metaphor	"La mappa indica il percorso che guida alla scoperta del tesoro" (*the*

[2] Reported meanings are those ones attested in Garzanti Dictionary.

		map shows the path to the treasure)
PILOTARE (synonyms = manovrare (*drive*))	Literal	"Andre Lagache e Rene Leonard guidarono per 1373 miglia la loro 3 litri" (*Andre Lagache and Rene Leonard drove their 3-litres car for 1373 miles*)

Reference dictionaries list 3 metaphorical meanings out of four. This imbalance has been verified even statistically: the analysis of about 800 occurrences, selected from a corpus drawn from the Italian online newspaper "La Repubblica", shows that in most part of the cases (about 73%) expectations of literality (subject: human and object: vehicle) are violated (fig. 3).

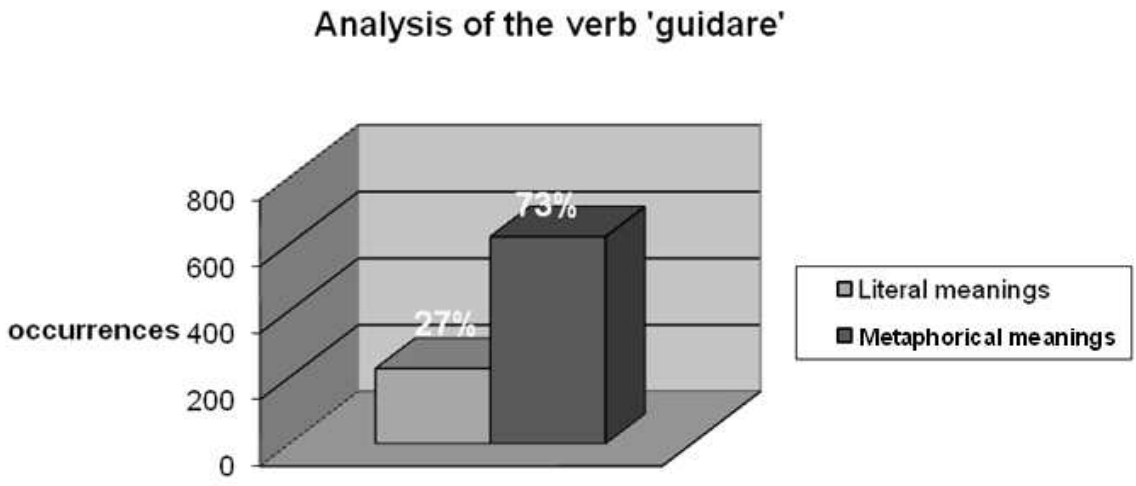

Figure 3. Representation of the ratio between metaforical and literal uses

In general, human beings seem to prefer to "guidare" (lit. *drive*) other humans organized in teams, departments, states. Sometimes subjects are non-human, but are instead intuitions, paths, rules. The variety of subjects and objects doesn't allow to establish either a rule or a fixed mapping. Variations, indeed, depend on cultural changes and are idiosyncratic. This is one of the reasons that suggested a dynamic approach based on a machine learning mechanism, able to build patterns on the fly so to adapt itself continuously as the corpora grows and changes.

7. Analogy-based processing of metaphors

This analysis shows how variability is strongly related to the way in which words are put together in sentences. In Fig.4 it can be seen how two different verbs such as "dirigere" (*lead*) and "guidare" (*drive*) which are close in virtue of a conceptual, metaphorical scheme, present differences as to the sentences in which they are involved for what concerns their co-occurring nouns ("guidare una squadra", lead a team; "guidare un'azienda", lead a company; "dirigere un'azienda", lead a company).

This inference process is called *extension* (Federici 1991) and accounts for both literal and non literal usages of natural language.

The proposed methodology allows us to identify patterns of similarity and to automatically create clusters of metaphorical uses, without need for much manual intervention. Finally, as a positive side effect, the application of this mechanism give

results (in terms of variability and regularity) that are intersubjectively verifiable and then more reliable with respect to analyses performed manually by linguists.

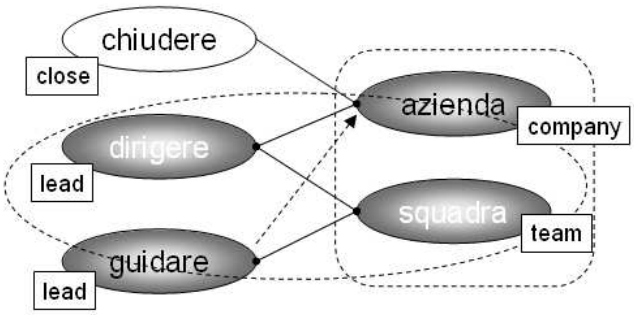

Figure 4. Analogical extension of object "azienda" (*company*) to verb "guidare" (*lead*).

In fig. 4, the knowledge contained in the Corpus (i.e. co-occurrences derived from corpus sentences) is represented by circles (nodes) connected by solid lines. Each pair of nodes connected by a solid line represents a co-occurrence attested in the Corpus. Each node contain (part of) the sentence, either the part that is shared among different sentences (e.g. the noun "squadra" (*team*) that is shared between the sentences "guidare la squadra" (lead the team) and "dirigere la squadra" (to manage the team)) or the part of the sentence that remains after isolating the shared part (e.g. the sequence "guidare" (*lead*) that appears in just one sentence "guidare una squadra" (to lead a team) when the shared part "squadra" (*team*) has been stripped off).

The nodes contained in the dotted area represent the *analogical paradigm* of the predicate "dirigere" (*manage*), namely the objects "azienda" (*company*) and "squadra" (*team*).

By structuring the knowledge contained in the corpus in this way, an automatic system can extend the usage of the word "azienda" (*company*) as an object of the predicate "guidare" (*lead*), that is it can extend the paradigm of the predicate "guidare" (represented in this example by its only object "squadra" (*team*)) with the new object "azienda" (*company*). The paradigm *extension* is represented by the dotted arrow that ends onto "guidare" and is triggered by the paradigms of the two *endings* (namely "guidare" and "azienda") that contain elements that recombine in at least one "fact" attested in the corpus (namely the sentence "dirigere la squadra" (manage the team)).

Paradigm extension can be rephrased as follows: if both "guidare" (*lead*) and "dirigere" (*manage*) can have "squadra" (*team*) as an object and "dirigere" (*manage*) can have "azienda" (*company*) as an object, than it is *justified* to suppose that "guidare" (*lead*), by *analogy* with "dirigere" (*manage*), can have "azienda" (*company*) as an object. The sentence "dirigere la squadra" (to manage the team), enclosed in the gray area, is the necessary *justification* to consider this inference as analogically motivated.

8. Conclusions

Many studies have shown the limits of conceptual approaches to the analysis of metaphors, especially because the adopted frameworks lack a satisfying empirical analysis. Instead, corpus-based metaphor analysis seeks to ascertain metaphor usage in larger text collections. The machine learning approach that we embrace in this paper relies on pattern recognition and analogical extension that, taken together, allow to overcome the boundaries between conceptual and lexical levels, grammar and use, language and thought, being able to evolve and grow along with corpora change.

References

Carbonell, J. G. (1982). Metaphor, An Inescapable Phenomenon in Natural Language Comprehension. Lehnert, W. G. and Ringle, M.H. (eds.) *Strategies for Natural Language Processing*, Lawrence Erlbaum Associate, Hillsdale, NJ, pp. 415-434.

Federici, S. (1991). SECS-NLC: A Self-Expandable Connectionist System for Natural Language Comprehension, in *Proceedings of AAAI Spring Symposium*.

Genette G., (1972). *La rhétorique restreinte*, in *Figures III*, Paris, Seuil, pp.21-40. Tr. It. *La retorica ristretta*, tr. it. In *Figure III. Discorso del racconto*, Torino, Einaudi, 1976, pp. 17-40.

Gola E. (2005). *Metafora e mente meccanica*, Cuec, Cagliari.

Gola E. (2008). "Metafore concettuali: che rapporto con il linguaggio?", in *Vie della metafora. Linguistica, filosofia, psicologia*, a cura di C. Casadio, primeVie edizioni, Sulmona.

Gola, E., Federici, S. (2000). Le regole informali del linguaggio naturale. Carapezza, M. and Lo Piparo, F., (eds) *La regola linguistica, Atti del VI Congresso di studi della Società di Filosofia del Linguaggio*, Palermo, Novecento.

Hardie, A., Koller, V., Rayson, P. and Semino, E. (2007). "Exploiting a semantic annotation tool for metaphor analysis". In: *Proceedings of the Corpus Linguistics 2007 conference*.

Koller, V., Hardie, A., Rayson, P. and Semino, E. (2008). "Using a semantic annotation tool for the analysis of metaphor in discourse". *Metaphorik.de* 15. http://www.metaphorik.de/15/.

Lakoff G., Johnson M. (1980). *Metaphors We Live by*, University of Chicago Press, Chicago (Illinois) and London (UK). (Trad. it. *Metafora e vita quotidiana*, Espresso Strumenti, Milano, 1982).

Lakoff G., Johnson M. (1999). *Philosophy in the Flesh. The Embodied Mind and its Challange to Western Thought*, Basic Books, New York.

Livingston, I. (2006). *Between Science and Literature: An Introduction to Autopoetics*. University of Illinois Press.

Martin J.H. (1990). *A Computational Model of Metaphor Interpretation*, Academic Press, New York.

Moravcsik J.M. (2002). Metaphor, Creative Understanding, and the Generative Lexicon. F. Bouillon, F. Busa (eds.), *The Language of Word Meaning*, Cambridge University Press, Cambridge (MA), pp.247-261.