Which way should we go?
Metaphoric expressions in lexical resources

Antonietta Alonge*, Margherita Castelli†

Sezione di Linguistica
Facoltà di Lettere e Filosofia - Università di Perugia
Piazza Morlacchi, 11
Perugia 06100 - ITALY

*antoalonge@libero.it, †castelli@unipg.it

Abstract

In this paper we address the issue of the encoding of information on metaphoric expressions in a (multilingual) lexical-semantic database for NLP applications. When analysing corpus data we find a huge number of metaphoric expressions which can be hardly dealt with by using as reference databases resources already developed. In particular, we have compared information contained both in dictionaries of Italian and in a WordNet-like resource – ItalWordNet – with actual uses of words found in a corpus. We thus put forward proposals to enrich a resource like IWN with relevant information.

1. Introduction

Lexical resources are a fundamental part of natural language processing systems and many research projects in the last decade have been devoted to build such resources, yet work is still needed to envisage what information should be encoded in which resource, and how.

In the context of the ISLE project[1] work is being carried out to design standards for multilingual lexicons. In particular, part of the research is devoted to identify and list the (maximal) set of *basic notions* which should be included in a multilingual lexical entry. Among the information to be encoded there are data such those contained in WordNet (WN) and WN-like resources (EuroWordNet – EWN – and ItalWordNet – IWN)[2], in which the meaning of words is described by means of various semantic/world-knowledge relations among word senses. However, in these resources (as in general in lexical resources) information on figurative, and in particular metaphoric, language is not systematically encoded. In general, the issue of how to treat these expressions has not been systematically dealt with within

research projects aimed at developing general/multilingual lexical resources for NLP (but see Fellbaum, 1998, for work related to WN; Nimb and Sandford Pedersen, 2000 for work carried out within the SIMPLE-EC project). Thus, aiming at devising a way to encode information on metaphoric expressions in a multilingual lexical entry, we have started by comparing information found both in IWN and in dictionaries of Italian[3] with data extracted from a large corpus of Italian built at the Istituto di Linguistica Computazionale, CNR, Pisa (Goggi et al., 2000).

In this paper we start by recalling the theory of metaphor proposed by Lakoff and Johnson (1980) and Lakoff (1993), that has spawned a variety of research in various fields connected with the study of language[4]. Then, we take into analysis examples of words displaying metaphorical sense extensions and discuss data from IWN (comparing it also with two dictionaries of Italian) and from the corpus. Finally, we propose a way of dealing with metaphors encoding in resources such as IWN.

2. Metaphor

Lakoff and Johnson (1980) and Lakoff (1993) show that metaphor is not just a poetical way of speaking, rather it is deeply embedded in our language, culture and the way we think. Metaphor affects how we experience and interact with the world and other people: "the human conceptual system is metaphorically structured and defined. Metaphors as linguistic expressions are possible precisely because there are metaphors in a person's conceptual system." (Lakoff and Johnson, 1980:6).

Within language structure we may distinguish i) a cognitive level (which organizes experience), ii) a conceptual (mental) level, iii) a semantic (linguistic) level, and iv) the level of the formal expression of concepts. At the 'cognitive level' human experience is categorized

---

[1] The ISLE project (International Standards for Language Engineering) is the prosecution of the EAGLES initiative (Expert Advisory group for Language Engineering Standards), which has seen successful development and broad deployment of a number of recommendations and *de facto* standards. ISLE aims at developing Human Language Technology standards within an international framework, in the context of the EU-US International Research Cooperation initiative (see http://lingue.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm).

[2] EWN was a project in the EC Language Engineering (LE-4003 and LE-8328) programme. Complete information on EWN can be found at its web site: http://www.hum.uva.nl/~ewn/gwa.htm. ItalWordNet has been developed within the SI-TAL (Integrated System for the Automatic Treatment of Language) Italian project, devoted to the creation of large linguistic resources and software tools for the Italian written and spoken language processing. The database was built by extending the Italian wordnet realized within EWN, and inherited from EWN its main characteristics (general structure, relations, etc.; see Alonge et al., 2000 and Roventini et al., forthcoming, for a detailed description of the IWN database).

[3] We have used "Il nuovo dizionario italiano Garzanti", 1991 and "Il nuovo Zingarelli", Zanichelli 1989.

[4] Browse the Center for the Cognitive Science of Metaphor Online for references at work on metaphor and links to other websites: http://philosophy.uoregon.edu/metaphor/metaphor.htm. See the Berkeley Conceptual Metaphor Home Page - database at http://cogsci.berkeley.edu/ for a database of conceptual metaphors.

according to the human mind structures. Some *image-schemas* (determined by our bodily experience) are used to build not only basic concepts, but also concepts which are not directly linked to physical experience. Thus, metaphoric linguistic expressions are manifestations of 'conceptual metaphors', i.e. metaphorical structures which are present in our mind and relate a concrete *source domain* with a more abstract *target domain*. In other words, metaphoric linguistic expressions are the superficial realization of the mapping we perform from a 'literal' source semantic domain to a 'figurative' target semantic domain. At least two consequences follow from this perspective which should be considered when building a lexicon:

i) metaphorical extension of word senses is a kind of regular polysemy (cf. Apresjan, 1980): e.g., *He arrived ('came here' or 'was born') when we were 20; He left us ('went away' or 'died') after some time*;
ii) generalizations govern inference models, i.e. those cases in which an inference model from a certain conceptual domain is used in another domain: e.g., *In our relationship we have faced many obstacles' → It has been difficult to go ahead.*

3. Metaphoric expressions in IWN and in a corpus of Italian

Sense distinctions vary widely across lexical resources.[5] Different dictionaries distinguish among different senses of words in a sort of arbitrary way since they are strongly influenced by the purpose of the resource (the target audience), and have different editorial philosophies with respect to 'lumping vs. splitting' of senses (Atkins, 1993; Kilgarriff, 1997). Dictionaries normally contain distinctions among 'literal' vs. 'figurative' meaning within a lexical entry. However, such information is in general, at best, 'incomplete':

i) information on metaphoric uses is not systematic in many sources, and different sources contain different information;
ii) potential metaphors are not encoded;
iii) when information on metaphorical sense extensions is present, there is generally no clear indication of the connection between the 'basic' and the 'extended' senses.

EWN first and IWN then were built using as source data dictionaries available in machine-readable form, thus they contain inconsistencies and shortage of data inherited

from dictionaries. While building first the Italian wordnet in EWN and then IWN, the problem of finding a coherent and principled way to deal with metaphoric extensions of word senses was always present but somehow neglected mainly because of the time limits of the projects. However, in order to (automatically) deal with real texts the issue has to be tackled identifying what and how should be encoded.

Consider, for instance, the verb a*rrivare* (to arrive): it has 9 senses in IWN, where the first one is a 'motion' sense and the others are all derived figurative senses with different degrees of proximity to the literal sense:

Synset	Definition
{arrivare 1, giungere 1, venire 2}	portarsi o essere portato in un luogo stabilito (to go or to be taken to a specific place)
{arrivare 2, giungere 4, raggiungere 2}	giungere ad un determinato punto o livello (di quantità, età, grado ecc.) (to reach a point, or a certain level – of quantity, age, degree, etc.)
{arrivare 3, venire 3, pervenire 1, …}	ottenere, conseguire un risultato, spec. per gradi e con difficoltà (to obtain a result, expecially gradually and with difficulty)
{arrivare 4, giungere 3, venire 9}	sopraggiungere, giungere con riferimento al procedere del tempo o nel tempo (to occur, to arise - with reference to the proceeding of time or in time)
{arrivare 5, venire 8, giungere 5, …}	assumere un nuovo stato, relazione o condizione (to enter a new state, relation, or condition)
{arrivare 6, andare 14, estendersi 2}	estendersi fino (to extend to)
{arrivare 7, giungere 8}	(fig.) giungere (figurative: to get to)
{arrivare 8, avvicinarsi 3}	diventare più vicino in senso temporale (to become closer in a temporal sense)
{arrivare 9}	affermarsi nella vita (to attain success)

Table 1: the senses of *arrivare* in IWN

Within the Garzanti dictionary we only find four senses, while within Zingarelli we find 6 senses. All the senses distinguished in the two dictionaries are also indicated in IWN. Thus, in this case, IWN seems rather rich of data on figurative senses. Nonetheless, by analysing corpus occurrences of *arrivare* we found that the various senses in IWN do not seem sufficient to account for all the occurrences of the verb. In the following we provide just two examples of senses from the corpus which we cannot easily trace back to the IWN senses:

- Celli, che alla letteratura *è arrivato* molti anni fa
  (Celli, who went into literature many years ago…)
- Lamberto Dini *è arrivato* alla politica dopo i sessant` anni
  (Lamberto Dini went into politics after he was 60)

- *sono arrivato* a spendere 400 mila lire di telefono al giorno
  (I went so far as to spend L. 400.000 for the telephone daily)

---

[5] Here we assume some sort of intuitive pre-theoretical notion of word-sense, which we are well aware that can be disputed. There has been much research devoted at the issue of what a word-sense is and if word-senses 'exist' at all and should be considered as the basic units of the lexicon. Although we agree with views according to which "word senses exist only relative to a task" (Kilgarriff, 1997: 1), and are at the same time appealed by proposals for 'coarse coding' (Harris, 1994)[5], we still believe that a WN-like structure, taking the concepts and the synsets referring to them as the 'building blocks' of the (mental) lexicon, is both appropriate as a representation of lexical knowledge (with the basic idea of a net linking the concepts) and can be used as a resource for NLP, provided that the possible uses and actual limits of such resource are kept clear.

- i giocatori *sono arrivati* a mordersi sul collo
(the players went so far as to bite each other's neck).

Both the senses involved here (which are actually very frequent in the corpus) could be linked to sense 7 of *arrivare* in IWN, since this sense is not clearly defined (is underspecified) in the resource. In fact, the two occurrences are quite different: although both could be seen as an instantiation of the general 'CHANGES ARE MOVEMENTS' metaphor (Lakoff, 1993), only the latter requires no *control* by an agent (we should rather speak of 'loss of control' by an agent) and always entails a result which is seen as negative. The difference between the two occurrences of *arrivare* becomes even clearer if we take into consideration the possible translations for them into English: again we use metaphoric expressions, but different expressions.

Besides being incomplete and inconsistent with respect to word senses which are very frequent in the corpus, the resources taken into consideration often lack information on infrequent or potential metaphoric sense extensions, which are also cases of regular sense extensions. A group of conventional conceptual metaphors which characterizes as a subset of the more general 'CHANGES ARE MOVEMENTS' metaphor is the following: 'BIRTH IS ARRIVAL', 'LIFE IS BEING PRESENT HERE', 'DEATH IS DEPARTURE'. Thus, we can say, for instance:

- Nostro figlio *è arrivato* (= è nato) dopo dieci anni di matrimonio.
(Our child arrived (= was born) ten years after our wedding)
- Lui non *è* più fra noi. (= è morto)
(He is not with us anymore. (= he is dead))
- Se ne *è andato* (è morto) all'età di venti anni.
He went away (he died) when he was twenty.

As we have seen, no such sense is encoded in IWN (or in the dictionaries considered) for *arrivare* (while we find encoded the senses indicated in the examples for *essere* and *andare*) even if this sense is attested (although infrequent) in the corpus:

- …di figli ne sono arrivati troppi.
(there arrived too many children).

If we then look for the senses provided for another verb which we may potentially expect to display the same regular sense extension of *andare* as *to die* – *lasciare* (to leave) –, we do not find any relevant information in our lexical resources as well, although also this verb metaphoric sense occurs once in our corpus:

- Mentre scrivo ci *ha* appena *lasciato*. La sua morte…
(While I'm writing he/she has just left us. His/her death…) .

In fact, these metaphoric uses of *arrivare* and *lasciare* are apparently rare, however they are not only possible but also quite normal in everyday spoken language (note that the corpus used is composed of texts taken from newspapers, magazines, essays, novels, etc.).

Potential or infrequent metaphoric word senses could be dealt with by confronting with another limit of IWN

and the dictionaries analysed, i.e. the lack of connections between basic and well-established metaphorical, extended senses. Let's consider an example of a frequently found metaphorical sense extension: the verb *separarsi* (to separate, to divide) has such a sense extension clearly encoded in sources. Within IWN we find, among the others, the following synsets containing the verb:

Synset	Definition
{separarsi 1, staccarsi 3, dividersi 2}	allontanarsi da qualcuno (to separate, to go far from someone)
{separarsi 4, lasciarsi 1, dividersi 4}	rompere un legame con qualcuno, specialmente di coppia (to separate, to break off a relationship with somebody, especially a love affair)

Table 2: senses of *separarsi* in IWN

Since enough information is encoded in IWN on the two senses of the verb, the database could be used to disambiguate it in the following sentences taken from the corpus:

- Elho sembrava leggermente a disagio all' idea di *separarsi* dagli altri, ma lui non vedeva l' ora di arrivare al fiume.
(Elho seemed to feel a bit uneasy when thinking to separate from the others, but he cannot wait to get to the river)
- egli avrebbe cercato di opporsi alla deportazione di suo fratello, da cui non voleva *separarsi*.
(he would have tried to set himself against his brother's deportation, since he did not want to separate from him)
- *Separarsi* dalla moglie, aggiunse, lo avvertiva un male…
(To separate from his wife, he said, seemed wrong).

The metaphor involved in this case is that according to which 'LOVE IS A JOURNEY' and the sense extension seen applies to a whole set of motion expressions which can be used to refer to love relationships. Thus, we could say that a man and a woman *have started walking together* (when they start a love relationship) and then that *their relationship has come at a turning point,* etc., using already attested metaphoric expressions, but also creating new ones. Neither in IWN nor in the dictionaries used there is indication of the connection of these two senses of *separarsi*, while it could be useful to have information on the existence of such a regularity of polysemy, in order to deal with novel metaphoric expressions involving verbs, or also other parts of speech, referring to the same basic conceptual domain.

4. Metaphors encoding

When comparing corpus occurrences of words with information encoded in IWN, or also in other lexical resources, one normally sees that there is a surprisingly high frequency of figurative senses in real texts and most of these senses are not described in such resources. The questions are: i) how should these figurative senses be

accounted for in a WN-like resource (in particular, in IWN)? And: ii) how should novel, potential uses of words be dealt with by referring to a resource such IWN?

As it is clear, IWN lacks precise information on very frequent metaphoric uses of words. By clustering corpus occurrences extracted from a general corpus of Italian it is possible to identify senses which could be added to the database to provide both a better account of a speaker's lexical knowledge and a set of data which are useful for various NLP tasks. Indeed, the data provided show that by analyzing a large general corpus various metaphoric expressions are clearly distinguishable which are not (consistently) identified in IWN or other resources. Since the necessity of adding corpora as sources for computational lexicons is probably unquestionable, our main point is that one should deal with these issues by adopting a well established and generally accepted theoretical framework like that proposed by Lakoff and Johnson (1980) and Lakoff (1993), within which a large system of conventional conceptual metaphors has been described. By adopting that perspective many subtle, but relevant, differences may be highlighted in a principled way. These should be encoded at the synset level to account for already well established word figurative senses. Of course, no lexical resource will probably ever be able to exhaustively account for the phenomenon which Cruse (1986) termed *modulation*, determining that 'a single sense can be modified in an unlimited number of ways for different contexts, each context emphasizing certain semantic traits, and obscuring and suppressing others" (Cruse, 1986: 52). However, each resource should be designed so to be as complete and coherent as possible.

What remains to be deepened is the issue of how to encode information on the systematic nature of conceptual metaphors. When we understand novel metaphoric expressions we make reference to a system of established mappings between concrete conceptual domains and abstract ones (e.g., the above mentioned mapping between the *journeys* domain and that of *love relationships*). That is, there is a pre-existent knowledge which constrains our possibility to produce and/or understand novel metaphoric expressions. In order to build a resource which actually accounts for our lexical-conceptual knowledge and can be used by as a resource for NLP, we have to find a way to encode also knowledge about mappings between conceptual domains resulting in potential metaphoric expressions production. This information should be encoded at a higher level than the synset level, since it is information on regular polysemy affecting whole conceptual domains.

In IWN, as in EWN, we have three fundamental levels of representation of semantic information:

   i) the synset level, where language-specific synsets information is encoded;

   ii) the level of the linking to the Interlingual-Index (ILI – an unstructured list of WN 1.5 synsets) to which synsets from the specific wordnet point in order to perform the linking between different language-specific wordnets;

   iii) the Top Ontology (TO), a hierarchy of language-independent concepts, reflecting fundamental semantic distinctions, which may (or may not) be lexicalised in various ways, or according to different patterns, in different languages: via the ILI, all the concepts in the

language specific wordnet are directly or indirectly (via hyponymy relations) linked to the TO.

Since the distinctions at the level of the TO are language independent, it is necessary to show metaphoric regular polysemy found in a specific language at a different level. Indeed, there are culture-constrained differences in the metaphor system (see, e.g., the differences linked to orientation reported by Lakoff and Johnson, 1980, determining for instance that in some cultures the *future is in front of us* and in others *the future is behind us*) which should receive a representation at some other level.

In EWN some cases of regular polysemy were dealt with at the level of the linking of each language-specific wordnet with the ILI. Via the ILI the generalizations over concepts were projected to the TO. Generalizations were stated directly at the level of the ILI and automatically inherited from all the synsets which in a language-specific wordnet were linked to the ILIs involved in the generalizations themselves. An automatically added generalization could be later manually deleted in case it did not apply to a specific language (cf. Peters et al., 1998). For instance, the lexeme *scuola* (school) in Italian has got (among others) two senses indicating one the *institution* and the other the *building*. This is a case of regular polysemy since many words indicating institutions also indicate buildings in Italian (as, of course, in other languages). Once we linked the school-institution and the school-building synsets to the appropriate synsets in the ILI, the system automatically added to both Italian synset another equivalence link, called EQ_METONYM, to a kind of 'composite ILI', clustering the 'institution' and 'building' ILIs into a coarser-grained sense group. Thus, our synsets, via the ILI, were linked to tops in the TO indicating concepts in different domains. A similar operation was automatically performed for senses reflecting diathesis alternations for verbs (related by EQ_DIATHESIS), such as causative and inchoative pairs. In case a kind of regular polysemy did not display in our language, we had to manually delete the automatically generated link to the relevant composite ILI.

In IWN the composite ILIs have not been used. However, we think that they could instead be adopted, by creating a much larger set of them, to account for regular metaphoric extensions of senses. In order to deal with culture-constrained differences in the metaphor system, instead of *a priori* identifying a set of composite ILIs to be automatically added to a language-specific wordnet (and eventually deleted, with various practical problems), it would be better to have the possibility to create for each language new composite ILIs which could eventually be shared among languages. Via the ILI links the connection between specific synsets in a language would also be shown at the TO level as connection (mapping) between concepts (linked to different conceptual domains). On the other hand, the mapping at the TO level could be used to infer which words might potentially display a certain metaphoric extension, when this is not encoded at the synset level. The link to a Top Concept is inherited along taxonomies, thus all the synsets directly or indirectly connected (through hyponymy) with another sysnet would inherit the links to Top Concepts related to different conceptual domains. Thus, even when specific information on a possible metaphoric sense extension of a

word is not encoded in the database it could be possible to derive it. Moreover, we are carrying out research to clarify which other relations (e.g., near_synonymy, xpos_near_synonymy, etc. – Alonge et.al, 1998) play a role in the definition of the subset of words which may (potentially) be involved by a certain regular sense extension.

5. Conclusions

In this paper we have discussed a way to deal with metaphoric expressions in WN-like databases which involves i) a more careful analysis of textual corpora and the reference to theoretical well-established views of metaphor to build computational lexicons; ii) the extensions of devices already used in EWN to encode information on the mapping between conceptual domains, causing the possibility to have certain metaphoric sense extensions in a language. The research on the latter issue needs to be refined; in particular, we want to deepen the issue of the weight that each relation has to determine when a concept belongs to a certain conceptual domain, for which metaphoric sense extensions apply.

6. References

Alonge, A., F. Bertagna, N. Calzolari, A. Roventini and A. Zampolli, 2000. Encoding Information on Adjectives in a Lexical-Semantic Net for Computational Applications. *Proceedings of the 1ˢᵗ Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, 42-50.

Apresijan J., 1980. Regular Polysemy. *Linguistics*, 142.

Atkins, B., 1993. Building a Lexicon: The Contribution of Lexicography. *International Journal of Lexicography*, 3: 167-204.

Cruse, D. A., 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Fellbaum, C., 1998. Towards a representation of idioms in WordNet. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal.

Goggi, S., L. Biagini, E. Picchi, R. Bindi, S. Rossi, R. Marinelli, 2000. *Italian Corpus Documentation*. LE-PAROLE, WP2.11.

Lakoff, G., 1993. The Contemporary Theory of Metaphor. In A. Ortony (ed.) *Metaphor and Thought. Second Edition*. New York: Cambridge University Press.

Lakoff, G. and M. Johnson, 1980. *Metaphors We Live by*. Chicago: University of Chicago Press.

Kilgarriff, A., 1997. 'I don't believe in word senses'. *Computers and the Humanities,* 31 (2): 91-113.

Nimb, S. and B. Sanford Pedersen, 2000. Treating Metaphoric Senses in a Danish Computational Lexicon – Different Cases of Regular Polysemy. *Proceedings of the 9ᵗʰ Euralex International Congress*.

Peters, W., I. Peters and P. Vossen, 1998. Automatic Sense Clustering in EuroWordNet. *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Granada, 409-23.

Roventini, A., A. Alonge, F. Bertagna, N. Calzolari, J. Cancila, R. Marinelli, A. Zampolli, B. Magnini, C. Girardi and M, Speranza, forthcoming. ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian. *Linguistica Computazionale*. Pisa: Giardini.