Encoding information on metaphoric expressions in WordNet-like resources*

Antonietta Alonge
Sezione di Linguistica
Facoltà di Lettere e Filosofia
Università di Perugia
Piazza Morlacchi, 11
Perugia 06100 - ITALY
antoalonge@libero.it

Margherita Castelli
Sezione di Linguistica
Facoltà di Lettere e Filosofia
Università di Perugia
Piazza Morlacchi, 11
Perugia 06100 - ITALY
castelli@unipg.it

Abstract

In this paper we address the issue of the encoding of information on metaphors in a WordNet-like database, i.e. the Italian wordnet in EuroWordNet (ItalWordNet). When analysing corpus data we find a huge number of metaphoric expressions which can be hardly dealt with by using as reference database ItalWordNet. In particular, we have compared information contained both in dictionaries of Italian and in ItalWordNet with actual uses of words found in a corpus. We thus put forward proposals to enrich a resource like ItalWordNet with relevant information.

1 Introduction

Lexical resources are a fundamental part of natural language processing systems and many research projects in the last decades have been devoted to build them. Notwithstanding, work is still needed to envisage what information should be encoded in which resource, and how, in order to deal with the central problem of word sense disambiguation.

As becomes clear when analysing textual corpora, novel uses of words occur frequently and, in particular, metaphoric new uses of them. Resources which simply list the different polysemous senses of a word without saying anything about how they relate to each other are not useful in such cases. One such resource is WordNet (WN), together with its 'relative-resource', i.e. EuroWordNet (EWN).[1]

In this paper we start by briefly recalling the theory of metaphor as a cognitive rather than a linguistic phenomenon, 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.[2] Then, we take into analysis examples of words displaying metaphorical sense extensions and discuss data from the Italian wordnet in EWN (ItalWordNet -

---

* The present paper is the outcome of a collaborative effort. For the specific concerns of the Italian Academy only, A. Alonge is responsible for sections 3 and 4; M. Castelli for sections 1, 2, and 5.

[1] EWN is a multilingual database developed within the homonymous project carried out in the EC Language Engineering programme (LE-4003 and LE-8328). Within the project, WordNet-like databases for various European languages were developed, connected by means of an Interlingual-Index. Complete information on EWN can be found at its website: http:// www.hum.uva.nl/~ewn/gwa.htm.

[2] Browse the 'Center for the Cognitive Science of Metaphor Online' for references at work on metaphor and links to other websites: http://philosphy.uoregon.edu/metaphor/metaphor.htm. See the Berkeley Conceptual Metaphor Home Page - db at http://cogsci.berkeley.edu/ for a database of conceptual metaphors. A similar database is being developed also for German and French: http://www.rrz.uni-hamburg.de/metaphern (cf. Eilts and Lönneker, 2002).

IWN).[3] Finally, we propose a way of dealing with metaphors encoding in resources such as EWN/IWN, which can be useful to obtain information which is necessary for word sense disambiguation.

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).

Thus, metaphor is a cognitive rather than a linguistic phenomenon, and word senses are related in terms of their underlying conceptual domains. Metaphoric linguistic expressions are manifestations of 'conceptual metaphors', i.e. metaphorical structures which are present in our minds and relate a concrete *source domain* with a more abstract *target domain*. In other words, metaphoric 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 for NLP applications:

i) metaphorical extension of word senses is a kind of regular polysemy (cf. Apresjan, 1973): 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 EWN-IWN

When looking for word sense distinctions within different lexical resources we see that these vary widely across resources.[4] 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 between 'literal' vs. 'figurative' meanings within a lexical entry. However, such information is in general, at best, 'incomplete':

1. information on metaphoric uses is not systematic in many sources, and different sources contain different information;
2. when information on metaphoric sense extensions is present, there is generally no clear indication of the connection between the 'basic' and the 'extended' senses;
3. data which could help to identify novel metaphoric expressions are not provided.

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 partly inherited from dictionaries, in particular with respect to figurative language. Consider, for instance, the verb *andare* (to go): it has 17 senses in IWN (reported below), two of which are strictly motion senses while the others are figurative senses somehow derived from the two basic ones, with different degrees of proximity to the literal senses.

[3] IWN was 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).

[4] We assume some sort of intuitive pre-theoretical notion of word-sense, which we are well aware that can be disputed. Much research has been devoted at the issues 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), 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 a resource are kept clear.

Synset	Definition
{andare 1, muovere 5, gire 1, ire 1}	muoversi, spostarsi, senza meta o senza che la meta sia indicata (to move or to change one's place without a goal, or without a specified goal)
{andare 2, recarsi 1, portarsi 1, trasferirsi 1}	muoversi, spostarsi verso una meta più o meno chiaramente definita (to move, to change one's place toward a more or less clearly defined goal)
{andare 3}	essere collocato, essere destinato ad essere messo in un dato luogo (to be located or to be intended to be put in a specific place)
{andare 4}	sentirsi in un certo modo (to feel in a certain way)
{andare 5}	trasformarsi (to transform - reflexive)
{andare 6, morire 1, dipartirsi 2, …}	cessare di vivere (to die, to cease living)
{andare 7, correre 12}	di moneta e simili, avere corso legale (of money, to be legal tender)
{andare 8}	dover essere (to have to be (done))
{andare 9, calzare 1}	essere adatto, calzare (to fit (s.o.))
{andare 10, piacere 1, garbare 1, …}	essere gradito, piacevole (to like)
{andare 11, precedere 2, progredire 2, …}	andare avanti, progredire in qualcosa (fig.) (to go ahead, to progress in something (figurative))
{andare 12, passare 10, dileguarsi 4, …}	venire meno, dileguarsi (to fade away, to disappear)
{andare 13}	continuare in un'azione (seguito da un gerundio) (to continue doing something)
{andare 14, estendersi 2, arrivare 6}	estendersi fino (to extend to)
{andare 15, dare 17, condurre 1, …}	dare accesso (to lead into)
{andare 16, funzionare 1}	adempiere la propria funzione (to work, to function)
{andare 17, muoversi 4, spostarsi 2}	(fig.) spostarsi (figurative – to move, to change one's opinion, etc.)

Table 1. Senses of *andare* in IWN

Senses 5, 6, 11, 12, 13, 14, 15 16 and 17 are clearly more directly derived from the two basic senses (either the first or the second): e.g., senses 5, 6, 11, 12 can all be linked to the general 'CHANGE IS MOTION' conceptual metaphor; sense 13 to the 'ACTION IS MOTION' metaphor, etc. The remaining senses seem also connected with the motion senses, although in a less direct way. [5] Only two of the metaphoric senses are marked as 'figurative' and no indication is provided of the connection between each metaphoric sense and the basic literal sense it is derived from. Moreover, if we take into consideration dictionaries of Italian like Zingarelli or Garzanti, [6] we find different sense definitions for *andare*: Zingarelli has 13 senses (with some information on connection of senses) and Garzanti has 11 (with no indication of sense connections). Finally, no information is provided, either in IWN or in the other resources, which could be used to automatically disambiguate novel metaphoric uses of the verb.

If we then look for occurrences of *andare* in a corpus of Italian (the PAROLE corpus, partly available at http://www.ilc.cnr.it/pisystem/demo/demo_dbt/demo_corpus/index.htm; cf. Goggi et al., 2000), we find occurrences of the verb which are hardly linked to the senses provided in our resources. Consider just two examples taken from this corpus:

a. Borg è già tornato e se n`è già *andato* in un mondo tutto suo (Borg has already come back and he has gone into his own world)
b. Altri sono *andati* con la memoria alle immagini televisive della guerra del Golfo (Others went with their memory to the television images of the Gulf war).

These two uses of the verb are quite frequent in spoken language, however they are not accounted for in the resources considered.

When comparing corpus occurrences of words with information encoded in IWN, as in other lexical resources, one normally sees that there is a surprisingly high frequency of figurative senses in

[5] E.g., sense 8 is found in sentences like "Questo lavoro va fatto subito" (This work has to be done immediately), where *andare* expresses a duty. We might suppose the existence of a conceptual metaphor like "TO FULFIL ONE'S DUTY IS TO MOVE TO A GOAL". This could be linked to a more general "ACCOMPLISHING SOMETHING IS REACHING A GOAL" metaphor, again connected with the "ACTION IS MOTION" metaphor. Of course, this analysis needs to be deepened; in particular, among other cases, one should also take into consideration the use of *venire* (to come), which in its basic sense indicates opposite direction with respect to *andare*, in sentences like "Questo lavoro *viene* fatto regolarmente" (This work is done regularly).

[6] *Lo Zingarelli 1999 in CD-ROM*, 1999, Zanichelli, Bologna; *Il nuovo dizionario italiano Garzanti*, 1991, Garzanti, Milano.

real texts but most of these senses are not described in such resources (cf. Nimb and Sanford Pedersen, 2000, for data identified within the SIMPLE EC project and the solutions proposed in that context). Alonge and Castelli (2002a) take into account corpus occurrences of the verb *colpire* (*to hit/to strike/to shoot*) and the noun *colpo* (*blow/stroke/shot*), and compare the results of this analysis with data found for these words within IWN, concluding that IWN lacks precise information on frequent metaphoric uses of *colpire* and *colpo*. 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 in other resources. Thus, how should these figurative senses be accounted for in a WN-like resource (in particular, in EWN/IWN)? Moreover, how should novel, potential uses of words be dealt with in a resource such IWN? We believe that the ability to cope with these issues cannot be set aside if IWN, or similar resources, has to be used for word sense disambiguation of 'real' texts.

4 Proposals for metaphors encoding in IWN/EWN

As already mentioned, by analyzing a large general corpus various well-established metaphoric expressions are clearly distinguishable which are not consistently encoded in IWN or in other resources. Since the necessity of adding corpora as sources for building 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 (cf. Alonge and Castelli, 2002a; 2002b). These should be encoded in EWN/IWN 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.

A more central issue to be tackled, however, is that of how to encode information on the systematic nature of conceptual metaphors, determining the possibility to produce and/or understand novel metaphoric uses of words. When we understand novel metaphoric expressions we make reference to a system of established mappings between concrete conceptual domains and abstract ones. That is, there is a pre-existent knowledge which constrains our possibility to produce and/or understand novel metaphoric expressions. For instance, 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' (cf. Lakoff and Johnson, 1980). Thus, we can say, for instance (examples are ours):

- 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.)

In IWN (or in the dictionaries considered) we find encoded the senses indicated in the examples for *essere* and *andare* but not for *arrivare*, even if this sense of the verb is attested (although infrequent) in the PAROLE corpus:

c. … 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:

d. 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*, although not frequent in our corpus (com-

posed of texts taken from newspapers, magazines, essays, novels, etc), are quite normal in everyday spoken language.

In order to build a resource which actually accounts for our lexical-conceptual knowledge and can be used 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:

- the synset level, where language-specific synsets information is encoded;
- 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 (by means of so-called 'equivalence-relations') in order to perform the linking between different language-specific wordnets;
- 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.[7]

The figure below exemplifies the relations among the three levels.

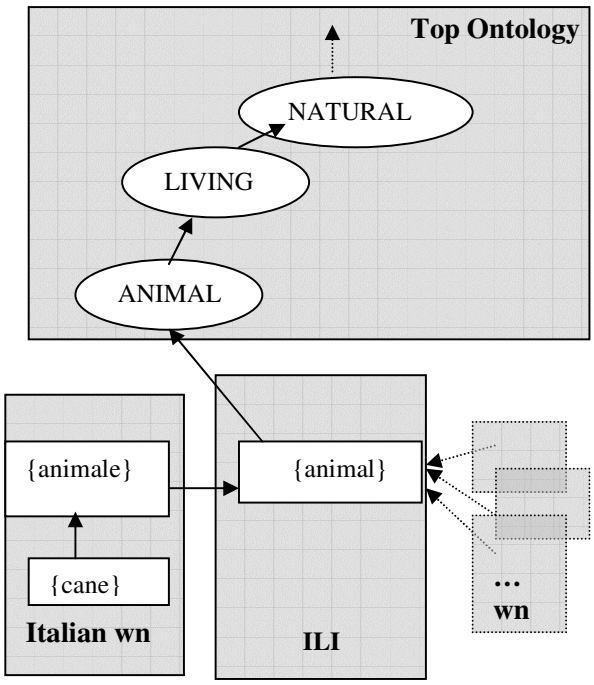

Figure 1. Levels of representation of semantic information in EWN/IWN

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 then 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 ILI synsets 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 related senses indicating

[7] Note that in EWN/IWN not all the synsets are directly linked to TO. Actually, only the so-called 'Base Concepts' (cf. Vossen, 1999) are explicitly linked to ILIs connected with Top Concepts. However, the links to Top Concepts are inherited by all the other synsets within the wordnets via hyponymy relations with the Base Concepts.

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 the Italian school-institution and the school-building synsets were linked to the appropriate synsets in the ILI, the system automatically added to both Italian synsets another equivalence link, called EQ_METONYM, to a kind of 'composite ILI unit', clustering the 'institution' and 'building' ILI synsets 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, the automatically generated link to the relevant composite ILI unit had to be manually deleted.

We think that an EQ_METAPHOR relation pointing to new composite ILI units should be created to account for regular metaphoric extensions of senses in EWN/IWN. Via the ILI links the connection between specific synsets in a language would also be shown at the TO level as connection (mapping) between top 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. Indeed, the link to a Top Concept is inherited along taxonomies within the language-specific wordnets, thus all the synsets directly or indirectly connected (through hyponymy) with another synset 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 would be possible to derive it. Consider the case of *lasciare* (to leave), mentioned above, and related conceptual metaphors. This verb has 9 senses in IWN (i.e., it is found within 9 synsets), one of which (sense 2) is defined as "andarsene da un luogo temporaneamente o definitivamente" (to go away from a place temporarily or definitively): this verb sense is a (direct) hyponym of the {partire 1, andare via 1} synset which, via an equivalence link to the {go, go away, depart, travel away} ILI synset, is connected with the Top Concepts 'BoundedEvent', indicating change of state; 'Location', indicating that the change referred to is a change of location; and 'Physical', indicating that the change of location involves 'physical entities'. As was done in EWN for other kinds of sense extensions, a 'composite ILI unit' should be created, clustering the 'departure' ILI synset (already linked to our {partire 1, andare 1} synset) and the 'death' ILI synset and accounting for the 'DEATH IS DEPARTURE' conceptual metaphor: then, the Italian synset already manually linked to the 'departure' ILI synset would also be connected, through an EQ_METAPHOR relation, to the 'death' ILI synset.[8] Consequently, the same synset would be, at the same time, connected both to the synset(s) indicating 'death' in Italian and to the relevant Top Concepts in the TO. All the hyponyms of {partire 1, andare via 1} would then inherit these connections; thus, also *lasciare* would display the same links, even if no specific information is encoded at the synset level. Again, cf. the figure below for a schematic representation of these relations.

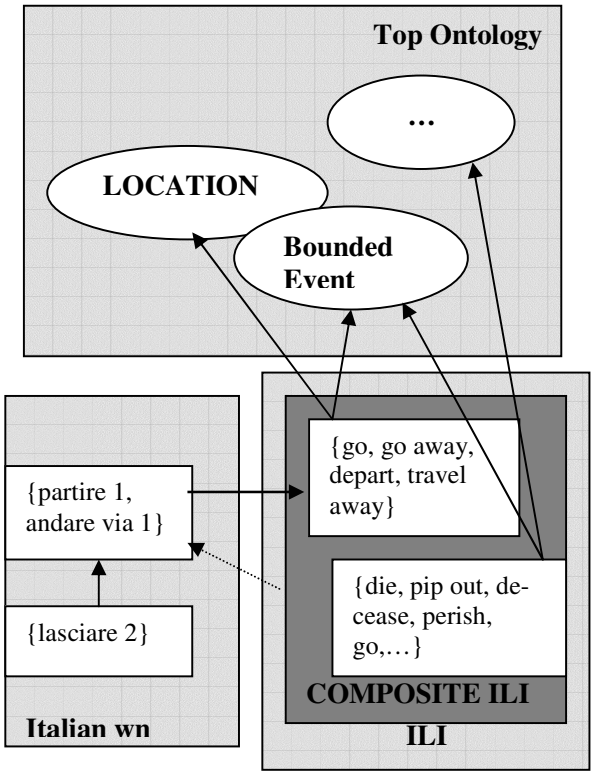

Figure 2. The *lasciare* links

[8] Note that for languages not displaying this sense extension the equivalence relation should be manually deleted.

In this way, information on potential metaphoric uses of *lasciare* (and, of course, other words in the wordnet) could be retrieved, by going from Top Concepts to ILI synsets and then to language-specific wordnets.

5 Conclusion

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 encode more detailed information at the synset level; ii) the extensions of devices already used in EWN in order 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 further carried out; in particular:

- we want to deepen the issue of the weight that other relations than hyponymy (cf. Alonge et al., 1998; Roventini et al., forthcoming) have in order to determine when a concept belongs to a certain conceptual domain for which metaphoric sense extensions inherited from the TO apply;
- we think that the EWN TO should be extended with more concepts, since at the moment it only contains very general semantic distinctions.

References

Alonge Antonietta, Francesca Bertagna, Nicoletta Calzolari, Adriana Roventini and Antonio 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.

Alonge Antonietta, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria Antonia Marti and Wim Peters. 1998. The Linguistic Design of the EuroWordNet Database. *Special Issue on EuroWordNet. Computers and the Humanities*, Vol. 32, Nos. 2-3: 91-115.

Alonge Antonietta and Margherita Castelli. 2002a. Metaphoric expressions: an analysis of data from a corpus and the ItalWordNet database, in *Proceedings of the 1st international WordNet Conference*, Mysore, India, 342-50.

Alonge Antonietta and Margherita Castelli. 2002b. Which way should we go? Metaphoric expressions in lexical resources, in *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain, 1948-53.

Apresjan Jurij D. 1973. Regular Polysemy. *Linguistics*, 142: 5-32.

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

Cruse D. Alan. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Eilts Carina and Birte Lönneker, 2002. The Hamburg Metaphor Database. *Metaphorik.de,* 3/2002.

Fellbaum Christiane. 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 Sara, Laura Biagini, Eugenio Picchi, Remo Bindi, Sergio Rossi and Rita Marinelli. 2000. *Italian Corpus Documentation*. LE-PAROLE, WP2.11.

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

Lakoff George and Mark Johnson. 1980. *Metaphors We Live by*. University of Chicago Press, Chicago.

Harris C. 1994. Coarse Coding and the Lexicon. In Fuchs C. and Victorri B. (eds.) *Continuity in Linguistic Semantics*, Benjamins, Amsterdam.

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

Nimb Sanni and Bolette Sandford Pedersen. 2000. Treating Metaphoric Senses in a Danish Computational Lexicon – Different Cases of Regular Polysemy. *Proceedings of the 9th Euralex International Congress*.

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

Roventini Adriana, Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Jessica Cancila, Rita Marinelli, Antonio Zampolli, Bernardo Magnini, Christian Girardi and Manuela Speranza. Forthcoming. ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian. *Linguistica Computazionale*. Pisa: Giardini.

Vossen Piek (ed.). 1999. EuroWordNet General Document. *The EWN CD-Rom* (see also: http://www.hum.uva.nl/~ewn).