Metaphors in Wordnets: from Theory to Practice

Antonietta Alonge

Sezione di Linguistica - Facoltà di Lettere e Filosofia University of Perugia Piazza Morlacchi, 11 Perugia 06100 - Italy anto.alonge@unipg.it

Birte Lönneker

Institute for Romance Languages University of Hamburg Von-Melle-Park 6 Hamburg 20146 - Germany birte.loenneker@uni-hamburg.de

Abstract

WordNet-like resources enhanced with a more systematic, consistent and efficient way of encoding data related to metaphors should improve the performance of a number of applications. As for the methodology to be followed, we stress two basic points: 1) the necessity of adding corpora as data sources and 2) the necessity of adopting a well-established and generally accepted theoretical framework like the one proposed by Lakoff and Johnson. The main part of this paper discusses how the practical work of adding information on metaphors to wordnets should be carried out. We propose a detailed analysis of one conceptual metaphor and its actual and possible representation within EWN. We also provide examples of how the results obtained could enhance the performance of applications using wordnets containing such information.

1. Introduction

Neither within the Princeton WordNet $(WN)^1$ nor within EuroWordNet $(EWN)^2$ the issue of how to treat metaphorical expressions was systematically dealt with. Alonge & Castelli (2002a) emphasized the limits of EWN in this respect, by analysing data encoded within the Italian wordnet (further developed as ItalWordNet – IWN, Roventini et al., 2003). They showed that in IWN:

- 1. information on metaphorical word senses is neither systematic nor consistent;
- 2. when information on metaphorical sense extensions is present, there is no indication of the connection between the 'basic' and the 'extended' senses;
- 3. data which could help to identify novel metaphorical expressions are not provided.

Thus, the authors made some proposals for the representation of metaphors in wordnets, while stressing at the same time the necessity to add corpora as further sources for wordnets and to analyse them by adopting as a reference framework the Cognitive theory of metaphor (Lakoff & Johnson, 1980). Other works (Alonge & Castelli, 2002b; Lönneker, 2003; Lönneker & Eilts, 2004) further deepened the issue and refined the proposals for a better encoding of data on metaphor in wordnets.³

In this paper we start by briefly recalling the Cognitive Metaphor (CM) theory. Then, we summarize the proposals already made towards a proper encoding of data on metaphors in wordnets. The main goal of this paper is however to discuss how the practical work of adding information on metaphors to wordnets should be carried out. We present a detailed analysis of one conceptual metaphor and its actual and possible representations within EWN. We also provide examples of how wordnets containing metaphor information could enhance performances of applications.

2. Cognitive Metaphor

Lakoff and Johnson (1980) showed that metaphor is deeply embedded in our language, culture and the way we think. Metaphorical linguistic expressions are of 'conceptual metaphors', manifestations i.e. metaphorical structures which are present in our minds and relate a (concrete) source conceptual domain with a (more abstract) *target* conceptual domain.⁴ The Cognitive Metaphor (CM) theory sheds light on at least the following intertwined aspects, which are relevant when building a lexicon for NLP applications:

i) metaphorical extension of word senses is a kind of regular polysemy: cf. e.g. *He arrived* (meaning 'came here' or 'was born') *when we were 20; He left us* (meaning 'went away' or 'died') *after some time*;

ii) existing conceptual metaphors are exploited/extended when new metaphorical meanings are created.

In general, successful new meanings are restricted to words belonging to source domains of conceptual metaphors. The domains and the mapping of domains are pre-existent knowledge which constrains the nature of successful novel metaphorical expressions. Moreover, conceptual metaphors may vary cross-culturally and when "a conceptual metaphor is universal, its universality obtains at a generic level, while the same conceptual metaphor shows cultural variation at a specific level" (Kövecses, 2002: 248). This has consequences with respect to the representation of information on metaphor within multilingual lexical resources.

3. More Data on Metaphors in Wordnets

An evaluation of current data in EWN against corpus data, analysed by referring to CM theory, allows to identify (conventionalized) metaphorical word senses missing from the database. This evaluation allows also to check:

¹ Http://www.cogsci.princeton.edu/~wn.

² Http://www.illc.uva.nl/EuroWordNet.

³ The issue was also discussed in a Panel session at the 2nd Global WordNet Conference: see http://www.fi.muni.cz/gwc2004.

⁴ A conceptual domain "is our conceptual representation, or knowledge, of any coherent segment of experience [...], such as the concepts of BUILDING or MOTION" (Kövecses, 2002: 247–248).

a) whether concepts found within the same (source) conceptual domain are properly linked among each other;
b) which concepts in the conceptual domain have (the expected) conventionalized metaphorical sense.⁵

Information on polysemy, currently not encoded for words displaying metaphorical sense extensions in EWN, could be added by means of a new *internal_relation*,⁶ e.g. DERIVED_FROM_LITERAL (Lönneker, 2003). This relation should be encoded between synset variants (i.e. synonyms within the synset), rather than between whole synsets, since it relates a specific word sense extension to the literal sense it derives from.

A harder issue to be tackled is the encoding of information on the pre-existent knowledge (i.e., knowledge about domain mappings), which constrains our possibility to produce and/or understand novel metaphoric expressions. As we think that this information could be useful for applications, we have to find a way to encode it. This information should be allocated at a higher level than the synset level, since it is information on regular polysemy affecting whole conceptual domains. Following Alonge & Castelli (2002b), we propose to draw generalizations on a language-independent level. These generalizations would be implemented as (literal-metaphorical) sense clusters encoded as coarse-grained sense groups at the level of the Inter-Lingual Index (ILI).⁷ In EWN these sense groups, or 'Composite ILI units' (CILIs), are already used for the representation of other cases of regular polysemy.⁸ Once a language-specific synset is linked to one of the ILIs in a CILI, the EWN database (Polaris system) automatically additional equivalence generates an relation ('EQ_METAPHOR') to the CILI itself. The equivalence relation could later be erased in case it does not apply to a specific language. That means that this relation is potential. The link to a CILI can be exploited to infer which synsets might have a metaphorical extension, when either this is not already conventionalized in a language or it is not represented due to shortcomings in the wordnet. Moreover, all the synsets in the same conceptual domain of a synset directly linked to a CILI are amenable to display a similar polysemy. That is, all the synsets linked to a (potentially) polysemous word by appropriate (potentially) internal_relations might display а metaphorical sense extension.9

4. From Theory to Practice

As a first practical result of the work carried out to identify metaphorical word senses within text corpora two metaphor databases were implemented and are constantly improved and updated: one in Hamburg (the 'Hamburg Metaphor Database' – HMD, cf. Lönneker & Eilts, 2004) containing French and German data,¹⁰ and one in Perugia, containing Italian data.¹¹ The data collected within these resources are also being used to evaluate information encoded within EWN.

In order to illustrate the various issues to be dealt with, we have chosen the Italian words *nascita* (birth) and *nascere* (to be born) for which some figurative senses are already encoded in IWN. We shall compare the encoded senses with the results of an analysis of corpus occurrences.¹² With respect to the noun *nascita* we notice that the three senses provided in IWN cover all the occurrences in the corpus. In particular, sense 2 (defined as "beginning, first period") expresses the conceptual metaphor CREATING IS BIRTHING,¹³ instantiated in the following corpus examples:

- è necessario favorire la nascita di un mercato
- (it is necessary to encourage the *birth* (beginning) of a market)
- la nascita di un Governo della Regione

(the birth of a Government of the Region)

This sense of *nascita* is an extension of the literal sense 1 of the noun (defined as "the event of being born"), but no present link may express such a relation in EWN. This information would become available by adding the proposed DERIVED_FROM_LITERAL relation. Moreover, we should create a CILI, clustering the literal and the metaphorical ILIs, and link it to both the literal and metaphorical synsets in IWN by means of an EQ_METAPHOR relation. The CILI could be used to encode the relation in other wordnets (in EWN). It would also link the language-specific synsets to different sets of Top Concepts.

Some weak points of the wordnet are detected for the senses encoded for the verb *nascere*, when compared with data from the corpus. Consider the following examples:

- dalla sua fatica è nato questo lungo studio
- (this study *was born* (derived) from his/her hard work) - nacque fra loro un'amicizia profonda
- (a deep friendship was born between them)
- da qui nacque la polemica
- (from this the argument *was born* (arose)).

We see that either an object, a state (relation), or an event may 'come into existence' (as a result of some other event), as can people (or living entities in general). Thus we should have one (and only one) sense of the verb *nascere* encoded in the wordnet to represent this

⁵ Note that "in metaphorical utilization, only some aspects of the source are utilized in metaphorical mapping, while the others remain unitilized" (Kövecses, 2002: 253).

⁶ 'Internal_relations' are relations between synsets or synset variants within a language-specific wordnet (Vossen, 1999).

⁷ The EWN ILI "is an unstructured list of meanings, mainly taken from WordNet1.5 [...]. The only purpose of the ILI is to mediate between the synsets of the language-specific wordnets." (Vossen, 1999: 8).

⁸ For instance, CILIs were created to cluster literal word senses and related metonymical ones (e. g., school as a 'building' and as an 'institution').

⁹ Via the ILI, clustered literal and metaphorical synsets are also linked to the relevant top concepts in the Top Ontology ("a hierarchy of language-independent concepts, reflecting important semantic distinctions" (Vossen, 1999: 8)). The connection between two related synsets in a language then reflects a mapping between different Top Concepts at the TO level. In terms of the CM theory, these different sets of Top Concepts represent (features of) the source and target domains.

¹⁰ http://www.rrz.uni-hamburg.de/metaphern/index_en.html

¹¹ The Italian database will be online by the end of 2004.

¹² We have used a subset of the PAROLE corpus of Italian, available at http://www.ilc.cnr.it/pisystem/demo_dbt/ demo_corpus/index.htm.

¹³ Actually, the label provided for the metaphor (cf. http://cogsci.berkeley.edu/MetaphorHome.html) is somewhat misleading, given that the example illustrating the metaphor is "This solution actually gives birth to a lot of new problems". Thus, a more appropriate definition would be 'CREATING IS GIVING BIRTH'. We shall, however, use the Berkeley label, since this is generally used also in other work.

metaphor. However, we find two synsets expressing the metaphor (one defined as "to be the outcome or consequence" and the other one as "to arise, to come up, of an event"), each containing various synset variants. If we try to replace *nascere* with each synset variant in the corpus examples, we may see that some synonyms are more specific than others (i.e., they are allowed in some contexts, but not in others). Thus, we might reorganize the data by creating one (more general) synset containing the metaphorical sense of *nascere* (together with some of the variants out of the existing synsets) and a more specific synset, encoded as a hyponym of the *nascere* synset, containing the more specific variants. Again, the relations proposed above should be encoded also for *nascere*.

As a further stage of our work, we check the current internal_relations encoded for the literal senses of *nascita* and *nascere* and evaluate their possible uses to 'populate' the domain of figurative word senses. E.g., *partorire* (to give birth, to deliver) is linked to *nascere* by a cause relation. Following the CM theory, since *nascere* has a derived metaphorical sense, also *partorire* should have a similar derived sense. Within the corpus we find for instance:

- È difficile partorire un nuovo spettacolo

(It is difficult *to give birth to* (to produce) a new show) This metaphorical sense of the verb is correctly encoded in the wordnet. However, the noun *parto* (childbirth, delivery), which is a *cross-pos-synonym* of *partorire*, has no metaphorical sense in IWN, although we frequently find this sense in the corpus; e.g.:

- il passaggio all'Unione Monetaria sarà un parto difficile (the passing to the Monetary Union will be a difficult *delivery* (achievement))

This sense of *parto*, and related links, should thus be encoded in the Italian wordnet.

An interesting case is that of *gestazione* (pregnancy, gestation), which is found in its literal sense in a synset (together with *gravidanza*) related to *nascere* by means of a (non-factive) cause relation. *Gestazione* has also a metaphorical sense encoded in the wordnet, and in fact we find examples of it in the corpus:

- Ignoravo che il decreto fosse in gestazione

(I ignored that the decree was *in gestation* (in progress)) The metaphorical synset does not contain the synonym *gravidanza*, for which no metaphorical usage is attested within the corpus. Thus, we should not add data at the synset level in this case. However, by linking the {*gestazione* 1, *gravidanza* 1} synset to the CILI clustering the literal and the metaphorical ILIs, we would ensure that novel metaphorical uses of *gravidanza* could be interpreted by applications using IWN. Indeed, by searching in the Internet we found the following example:¹⁴

- Una piacevole, impegnativa e interessante gravidanza, iniziata due anni fa, ha avuto come esito questo libro

(A pleasant, engaging and interesting *pregnancy*, started two years ago, has had this book as a result)

The fact that metaphorical extensions of variants inside the same synset can be more or less conventionalized is in line with results of other corpus analyses. For example, Małgorzata Fabiszak (Poznan) found the English word *battleground* (illustrating the ARGUMENT IS WAR metaphor) to show 14 literal and 77 figurative uses in 102 hits from the BNC, while an analysis of its synonym *battlefield* yielded 64 literal and 28 figurative uses in 100 hits.¹⁵ Fulgosi & Tuđman-Vuković (2000) found the Croatian word *korijen* (root) to have a metaphorical meaning (illustrating PEOPLE ARE PLANTS, IDEAS ARE PLANTS) in 68% of its 476 occurrences in the Croatian National Corpus. Its synonym *korijenje* occurs metaphorically only in 24% of the corpus hits.

Still another case in the Italian data is that of *abortire* (to abort), linked to *nascere* by means of a cause (negative) relation. This verb has an established metaphorical sense in Italian, as indicated by the corpus example given below, but no metaphorical sense encoded in the wordnet: - sepolto in compagnia dell'altro progetto abortito

(buried together with the other project that was aborted) In this case, the appropriate information should be encoded at the synset level.

The adoption of the CM perspective led us to reconsider also various synsets which are not properly linked in IWN to those analysed so far. Within the source domain of CREATING IS BIRTHING we find e.g. padre (father), madre (mother), madrina (godmother), figlio (child), concepire (to become pregnant), culla (cradle), etc. The analysis of the corpus data shows that most of these words already have an established metaphorical sense, but some don't: e.g., we find padre della fisica moderna (father of modern physics), but no similar expression for *madre*.¹⁶ IWN correctly represents these data only in some cases. Thus, for all these words we should add the relevant data and relations. By encoding a DERIVED_FROM_LITERAL relation only for those metaphorical word senses attested in a reference corpus and an EQ_METAPHOR relation for all the synsets in the source domain under analysis we ensure that applications can distinguish between different levels of conventionalization of metaphorical usage of words.

The detected shortcomings are not limited to IWN. In HMD, data on the CREATING IS BIRTHING metaphor were found to occur in corpora centered on political events. A database query for the target domain CREATING and the source domain BIRTHING yields twelve results,¹⁷ eight for French and four for German. The figuratively used words are not limited to *birth* and corresponding verbs. For French, we obtain four verbs or verbal phrases (naître 'to be born', voir le jour 'to "see the day"; to live; to appear', accoucher 'to give birth', baptiser 'to baptise'); the French wordnet does not provide the metaphorical meaning for any of these. In fact, only *naître* is present in EWN, but in a literal meaning. Two French nouns (naissance 'birth' and père 'father'), which both occur twice in HMD, can be found in EWN; however, the metaphorical meaning is encoded only for naissance. As for the four German lexemes, two of them (Geburtsstunde 'hour of birth', Taufe 'baptism') are covered by the German part of EWN, but only in their literal meaning.

¹⁵ Personal communication on 28 February 2004.

¹⁶ And, actually, *madre* does not seem to have such an established metaphorical use. However, we may not exclude, in principle, a possible similar use of this word.

¹⁷ Query performed on 26 February 2004.

¹⁴ Cf. http://www.mag4.it/prin/giovalibr.htm.

5. Some Consequences for Applications

One can think of many applications that would profit from a more systematic representation of metaphors in wordnets, both for direct use of humans (e.g. in language learning) and for machines, and both on the understanding and the generation side. For illustration, we give a more detailed explanation of how metaphors affect the specific NLP subtask of automated sense clustering. This task consists in finding out which senses of a sense inventory stemming for example from WN - are similar and can therefore be grouped. Algorithms for Word Sense Disambiguation and related tasks usually perform better for "coarse-grained senses" corresponding to the clusters. This is so because less, and at the same time more distinct, senses have to be distinguished. However, on which grounds senses are to be regarded as similar depends partly on the application (Information Retrieval, Question Answering, Machine Translation...) that will use them (cf. Chugur, Gonzalo, & Verdejo, 2000; 2002), and so does the algorithm to produce the clusters.

Several experiments have been performed using a crosslinguistic measure of sense similarity, which assumes that two word senses are similar if they are often lexicalized in the same way in a set of languages (cf. Resnik & Yarowsky, 2000). Chugur, Gonzalo, & Verdejo, (2002) found that the result quality of this similarity measure highly depends on the nature of the relation between word senses. In particular, it does not yield good results for meanings related by metaphor. Examples pointed out by Chugur, Gonzalo, & Verdejo, (2000; 2002) are blind ('unable to see' vs 'irrational') and steering ('act of steering and holding the course' vs 'guiding, guidance'). An IR application would actually perform better if these sense pairs were not clustered, because they "belong to different semantic fields and, consequently, tend to occur in distinct documents" (Chugur, Gonzalo, & Verdejo, 2000). Agirre & Lopez de Lacalle (2003) compare different similarity measures including those based on contexts (sentences, sentence snippets) showing the polysemous word in only one of its senses. Automated clustering based on contextual measures was found to perform better (as compared to human judgement) than clustering based on cross-linguistic similarity. An "explicit account for metaphors" in wordnets as already called for by Chugur, Gonzalo, & Verdejo, (2002:39) would thus help to correct, evaluate and interpret the results of similarity measures.

6. Conclusions and Future Work

The identification of literal and related metaphorical ILIs in EWN is well underway. The resulting CILI groupings will be made available on the Global WordNet Association website,¹⁸ in order to reach agreement on them among researchers involved in wordnet building. They should then be integrated into the ILI. The clusters produced at this level would mainly be used to derive (potential) metaphorical word senses in the languagespecific wordnets along the lines illustrated above. In parallel, we hope to be able to propose more elaborate systems of both conceptual domains and EWN TO features based on the work we are carrying out to develop the French/German and Italian metaphor databases. In a further step, the metaphor databases will be used to make complete proposals of concrete data for enhancing the metaphor representation in EWN, e.g. in the form of an add-on or patch that EWN users can integrate into the existing data.

References

- Agirre, E. & Lopez de Lacalle, O. (2003). Clustering WordNet Word Senses. In Proceedings of the Conference on Recent Advances on NLP (RANLP'03) (pp. 11–18). Borovets, Bulgaria.
- Alonge, A. & Castelli, M. (2002a). Metaphoric Expressions: an Analysis of Data from a Corpus and the ItalWordNet Database. In Proceedings of the 1st International WordNet Conference (pp. 342–350). Mysore, India.
- Alonge, A. & Castelli, M. (2002b). Which Way should We Go? Metaphoric Expressions in Lexical Resources. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (pp. 1948–1953). Las Palmas, Canary Islands, Spain.
- Chugur, I., Gonzalo, J. & Verdejo, F. (2000). A Study of Sense Clustering Criteria for Information Retrieval Applications. In Proceedings of the OntoLex-2000 Workshop (pp. 25–42). Sozopol, Bulgaria.
- Chugur, I., Gonzalo, J. & Verdejo, F. (2002). A Study of Polysemy and Sense Proximity in the Senseval-2 Test Suite. In Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions (pp. 32–39). Philadelphia, PA, USA.
- Fulgosi, S. & Tuđman-Vuković, N. (2000). Metaforična značenja i korpus; neke implikacije za leksikografiju.
 In D. Stolac, N. Ivanetić, & B. Pritchard (Eds.), Primijenjena lingvistika u Hrvatskoj (pp. 137–148). Zagreb.
- Kövecses, Z. (2002). Metaphor. A Practical Introduction. New York: OUP.
- Lakoff, G. & Johnson, M. (1980). Metaphors We Live by. Chicago: University of Chicago Press.
- Lönneker, B. (2003). Is there a Way to Represent Metaphors in WordNets? Insights from the Hamburg Metaphor Database. In Proceedings of the Workshop on The Lexicon and Figurative Language, 41st Annual Meeting of the ACL (pp. 18–26). Sapporo, Japan.
- Lönneker, B. & Eilts, C. (2004). A Current Resource and Future Perspectives for Enriching WordNets with Metaphor Information. In Proceedings of the 2nd International WordNet Conference (pp. 157–162). Brno, Czech Republic.
- Resnik, P. & Yarowsky, D. (2000). Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. Natural Language Engineering 5(2), 113–133.
- Roventini A., Alonge A., Bertagna F., Calzolari N., Cancila J., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A. (2003). ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian, in A. Zampolli, N. Calzolari, L. Cignoni (Eds.), Computational Linguistics in Pisa, Special Issue, Vol. XVI-XVII, Pisa, Giardini, 745–792.
- Vossen, P. (1999) (Ed.). EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document (available at: http://www.illc.uva.nl/ EuroWordNet/docs.html).

¹⁸ http://www.globalwordnet.org/