Integrating Two Semantic Lexicons, SIMPLE And ItalWordNet: What Can We Gain?

Adriana Roventini, Marisa Ulivieri, Nicoletta Calzolari
Istituto di Linguistica Computazionale, CNR
Via Moruzzi 1 – Pisa - Italy
adriana.roventini@ilc.cnr.it, marisa.ulivieri@ilc.cnr.it, glottolo@ilc.cnr.it

Abstract

In the last years, at the Institute for Computational Linguistics in Pisa, a few lexical resources have been developed aiming at encoding complex lexical semantic information. ItalWordNet and SIMPLE are two of these resources which, tackling semantics in the lexicon from different points of view, and being at least partially complementary, could certainly profit from linking each other. These resources in fact evidence different aspects of the lexical information: in SIMPLE, which adds a semantic layer to the morphological and syntactic ones developed in PAROLE, the connections between semantics and syntax are preeminent; ItalWordNet (as the Princeton WordNet and then EuroWordNet) is built around the basic notion of a synset and various semantic relations are encoded between synsets while syntactic aspects are not taken into consideration. In the paper we describe an experiment we carried out, aimed at exploring the feasibility of linking these lexical resources, being convinced that a noteworthy gain could be achieved through this operation. As we will show in the following some problems came in the foreground but also considerable advantages concerning the coherence and the completeness of both of them.

1. Introduction

Even though we speak a lot about the need of integrating different resources, components, tools, etc., the reality is often different as many resources are, for contingent facts, developed independently. This is true also of the SIMPLE and ItalWordNet semantic lexicons, built in our institute in two different European and then National projects, with the usual time constraints which made it impossible for the two teams to work together. It is however absurd to keep the two resources separate, because each one has obviously much to gain from an integration, given the complementarity of the two lexical models and of the methodology of building the two resources.

ItalWordNet (IWN) was first developed within the EuroWordNet (EWN)[1] project (Vossen, 1999) and then extended in the framework of an Italian national project for the automatic treatment of the language SI-TAL[2]. IWN (Roventini et al. 2002, forthcoming) is a large lexical-semantic database containing semantic information for about 50,000 synsets of nouns, verbs, adjectives, adverbs, and a subset of proper nouns. The information is encoded in the form of lexical-semantic relations between pairs of synsets (synonym sets). The most important relations encoded, often using as sources machine-readable dictionaries, are synonymy and hyponymy, however a rich linguistic model was designed containing many other lexical-semantic relations which are encoded for various subsets of Italian nouns, verbs and adjectives. All the synsets are also linked to WordNet 1.5, the Princeton Wordnet database (Miller et al. 1990).

The SIMPLE semantic lexicon has been developed in the framework of a European project[3] aimed at building core multipurpose and harmonised computational semantic lexicons for 12 European languages, linked to the morphological and syntactic ones built during the PAROLE[4] project. These lexicons consist of 10,000 semantic units (SemU) of nouns, verbs and adjectives. Besides the model and the encoding criteria, the 12 SIMPLE lexicons share a core of senses selected among the EWN "base concepts". This was conceived as both as the beginning of a multilingual linking among the SIMPLE lexicons and as the starting point for an eventual connection between the two resources, the SIMPLE lexicon and the IWN lexical-semantic database. The SIMPLE Italian lexicon is now being extended within the national project "Corpora and Lexicons of Italian written and spoken language (CLIPS)" (Ruimy et al. 2002). The theoretical model underlying this lexicon is based on the EAGLES recommendations (Sanfilippo et al. 1999) and on a revised version of Pustejovsky's Generative Lexicon (Pustejovsky, 1995).

In this paper we report about the first results of a feasibility study for the linking - to be done at least partially automatically - between the two lexicons.

2. Linking advantages

By linking these two semantic lexicons, SIMPLE and IWN, we can obtain two orders of advantages for both of them.

The first advantage is obviously the main goal of the linking, i.e. the possibility of using the two lexicons together, thus enriching each one of the information types characteristic of the other. IWN can benefit by the syntactic information encoded in SIMPLE, thus gaining

[1] EWN was a project in the EC Language Engineering (LE4003) programme. Complete information on EWN can be found at its web site: http://www.hum/uva.nl/~ewn.

[2] The SI-TAL project : 'Integrated Systems for the Automatic Treatment of Language' was a National Project, coordinated by A. Zampolli, devoted to the creation of large linguistic resources and software tools for the Italian written and spoken language processing. Besides IWN, within the project were developed: a treebank with a three level syntactic and semantic annotation, a system for integrating NL processors in applications for managing grammatical resources, a dialogue annotated corpus for applications of advanced vocal interfaces, software and tools for advanced vocal interfaces.

[3] SIMPLE was the EC project (LE-8346).

[4] The project *PAROLE* ('Preparatory Action for Linguistic Resources Organisation for Language Engineering') was funded by the European Community (1996-1998).

the rich syntactic and semantic subcategorisation, by the extensive domain encoding, qualia relations, etc., while SIMPLE could take advantage by the extensively encoded synonymy and taxonomy relations of IWN. Furthermore, another advantage for SIMPLE will be the possibility of being put in relation with WordNet 1.5 through the IWN mapping, and through it to the other EWN lexical-semantic databases for at least eight European languages, thus gaining a multilingual dimension. Another important indirect gain is the fact that the Italian TreeBank (Corazzari et al. 2001, Calzolari et al. 2001) has been semantically annotated with reference to IWN. After the linking, it will be automatically annotated also with respect to SIMPLE, and to its semantic types.

In addition to these evident gains, which are the main reasons of this linking operation, once the two lexicons are integrated, another, not minor, advantage is already obtained during the linking process, i.e. the achievement of a much greater coherence and consistency of both lexicons. The linking process can in fact be considered as a sort of reciprocal evaluation of the two lexicons. This is particularly important for a semantic lexicon, where it is practically impossible to avoid subjectivity in the classifications, despite the availability of criteria for assigning categories in both lexicon specifications.

Moreover, being SIMPLE much smaller than IWN, we can use the synonymical relations encoded in synsets to facilitate a quicker encoding of variants not yet encoded in SIMPLE: they will share the Semantic Type, and consequently most of the template information, thus speeding up the encoding process and ensuring coherency.

2.1 Diverging aspects

There are a few aspects where the two lexicons diverge. The most important are the following:

- they have different top ontologies – even though partially mappable: SIMPLE has semantic types organised in a hierarchy, with associated templates of information providing the semantics of the type, while IWN has a set of rather flat top semantic features (just labels);
- the basic unit to which all the information is related in SIMPLE is the Semantic Unit (SemU), while in IWN it is the Synset. This last difference has important consequences in a multilingual environment, e.g. for machine translation. It is in fact not always the case that the variants in the synset are interchangeable translations in any context.

Despite the differences, we can still exploit the partial mappability of the ontological information for automating the linking. We can even turn the mismatch in the basic units into an advantage (as we see in the following), because it can be usefully exploited to improve coherency and to facilitate encoding of new senses.

3. The experiment

In our experiment, we took into consideration both first order and second order entities. In particular, for the first order, a few nouns belonging to the semantic domains of F*ood* and B*uilding*, and for the second order a set of verbs and a few nouns belonging to the *Feeling* domain. The two classes of concrete nouns do not present many problems, and the SIMPLE SemU, as expected, appeared forwardly linkable to the IWN synsets. Obviously even concrete nouns, in a few cases, show different encoding, but usually both resources connect these classes to the same ontological concepts and this makes it possible to get an automatic linking between them.

The same is not true as far as the second order entities are concerned. These sometimes refer to different ontological concepts in the two resources, and are more complex from many points of view. In any case, the fact that with abstract senses we find more often discrepancies in semantic type assignment is a sign of the difficulty of providing explicit and discriminating classification criteria, and consequently of the inevitable subjectivity in semantic classifications. Cases of discrepancy can become however a useful hint of the more problematic areas in the ontology, thus forcing a more careful analysis of these areas.

3.1 First order entities

Let us consider a few examples of different mapping situations between the two resources. If we consider a concrete noun belonging to the *Food* domain as *crostata* (tart), we do not find any problem for an automatic link. In both the resources we find the same ontological pattern and a similar coding. In this case IWN could gain further information because SIMPLE shows, for the food domain, a richer encoding, compared with IWN, through the qualia structure. The considered SemU *crostata* (tart) is related, for example, with the verb *impastare* (to knead) and with the noun *cottura* (baking) by the semantic relation SRCreatedby, and with the verb *mangiare* (to eat) by the semantic relation SRObjectoftheactivity. These semantic relations are specific instances of qualia.

Nevertheless, if we consider more complex concepts such as for example that expressed by the word *casa* (house), which presents in both the resources a rich coding, we can find, besides reciprocal gains in information, some coherence problems.

As regards this concept, IWN and SIMPLE have in common the ontological pattern: object – building – artifact, and a few semantic relations: one Hasaspart relation with *vano* (room) and one Usedfor relation with *abitare* (to live in), even if in IWN this telic feature is represented by the more punctual relation of Role_Location. Furthermore there are a few different relations which could result in a reciprocal advantage. For example, in IWN we also find a Role_Target_Direction which establishes a link between *casa* and the verb *rincasare* (to return home), and in SIMPLE a SRPolysemyHumanGroup-Building relation which creates an explicit link between two different senses of *casa*, showing this typical sense shifting or regular polysemy. In IWN the regularity of this sense shifting is not evidenced in the coding. This fact constitutes a good example of advantages (above) but of problems at the same time, because in IWN *casa* is a member in the synset {*casa*, *abitazione*, *dimora*} and the sense shifting does not apply to *abitazione* and *dimora*. So, for this kind of relation, the fundamental difference existing between the two resources (i.e. to be created around the different basic notions of SemU and synset) can play a negative role in the automatic linking.

Considering the wider meaning of *casa* compared with the other members of the synset, we could overcome this problem making a change in IWN by encoding, *abitazione* and *dimora* as Near_Synonym of *casa*. Otherwise – and this will probably be the preferred solution - the linking should be established between a SIMPLE SemU and a IWN variant in a synset.

As another example of first order entities, if we consider the coding of the word *coniglio* (rabbit) we see that in IWN there are only two word meaning encoded: *coniglio* as "mammal rodent" and the metaphoric use of *coniglio* to mean "timid person". The meaning of "food" in IWN can be deduced from the definition but it is not explicitly encoded, and the last sense of "fur" does not appear in any way. In SIMPLE/CLIPS four semantic units have been created, respectively, for the animal, person, food and material (i.e. fur) word senses. So these two last senses could be acquired from SIMPLE where both the SRPolysemy animal-food and the SRPolysemy animal-material are encoded.

Moreover, in SIMPLE, the use of appropriate qualia relations allows to link together all these semantic units, specifying the kind of relationships between them. Taking the animal-typed semantic unit as a keyword and accessing, via queries through the SIMPLE database, all qualia relations in which it is used, all the connected entries can be retrieved: the relation ' metaphor' encodes the link to *coniglio* as a timid human, two polysemic relations account for the typical regular polysemy between animals and their meat as well as between animals and the leather derived from them, a constitutive relation links the animal to the location in which it is grown *conigliera* (rubbit hutch), etc. In this way in SIMPLE/CLIPS the encoding of entries regarding e.g. animal entities makes it possible to retrieve, via a keyword, all information concerning animal' s world. In this specific case, the link between the two lexical resources implies, as far as IWN is concerned, two advantages: it is possible, on one hand, to enrich the synsets and, on the other hand, to capture all the connected information.

3.2 Second order entities

A first example concerns an abstract noun belonging to the feeling domain: *odio* (hate) which in IWN shows the ontological pattern "mental, experience, dynamic" and in SIMPLE PsychologicalEvent-Agentive. The two ontological typings are compatible and mappable with each other. Moreover both the resources put in relation the noun with the verb *odiare* (to hate) using different but equivalent relations: IWN the XPOS_Near_Synonymy and SIMPLE a derivation relation. The difference in this case is in sense granularity and it is due to a more precise sense differentiation in IWN between *odio* as strong ostility towards a person and *odio* as intolerance towards something (as in the sentences: I hate snakes / jewellery / mushrooms) which in SIMPLE are unified in only one word sense. The same type of discrepancy verifies with the word *invidia* (envy): one sense for SIMPLE and two senses for IWN. Futhermore *invidia*, in the first word sense, is joint to *gelosia* (jealousy) with the meaning of spiteful, malign feeling, and in the second one is defined as feeling of sincere admiration. Other cases observed, as for example *gelosia* (jealousy) show also different ontological patterns and still different sense differentiations.

We also focused our attention on a set of verbal entries, which has been the real test bench for our experiment, analysing the possibilities of linking for about 100 verbs entries. These verbs, randomly selected, belong to many different classes. A few examples can illustrate the work and the preliminary results.

The verb *macchinare* (to plot) in IWN shows the ontological path "mental-purpose-agentive" and is a variant in the synset {*macchinare*, *tramare, tessere, ordire, architettare*}. In SIMPLE *macchinare* has no synonyms, and furthermore has a different ontological path, "Purpose-Act", which does not represent the mental process evidenced in IWN. In this case the IWN representation has been preferred and the SIMPLE verb has been changed and encoded as psychological event.

Another clarifying example concerns one sense of the verb *attaccare* (to attack) which in IWN is represented by the synset {*attaccare, oppugnare, aggredire, assaltare, assalire*}. In SIMPLE we found this sense codified as "Purpose-Act", *aggredire* as "Relational-Act", and *assaltare* and *assalire* as "Cause-Motion". Also in this case the comparison evidenced a lack of coherence in the SIMPLE classification, which has been corrected by moving all these senses under the "Purpose-Act" semantic type. The coding in SIMPLE, in a few cases, turns to be less coherent for the reason that synonymical words have been encoded separately, without bearing in mind the specific phenomenon of synonymy.

Another example is constituted by the IWN synset {*intendersi, capirsi, andare d'accordo*} (to get along with s.o.). This synset can be mapped to the two first different Semantic Units in SIMPLE given that multiword expressions are not yet encoded in SIMPLE. But the ontological pattern is different, because these verbs in SIMPLE are considered PsychologicalEvent and have an isa relation with *evento* (event), while the IWN synset shows the ontological pattern "social, property, mental, communication" and the hyperonym is the stative *essere* (to be) which is related to the ontology feature "property".

Another case is the IWN synset {*andare a male, avariarsi, rovinarsi, guastarsi, avariare, deperire* } (to go bad, to perish) which has as ontological pattern "bounded event" and as hyperonym {*trasformarsi, divenire, diventare, farsi*} (to become). The synset should be linked to the corresponding Semantic Units in SIMPLE, but the path is very tortuous because in this resource *guastare* has hyperonym *danneggiare* which has hyperonym *rovinare* which has hyperonym *cambiare* (to change) while *avariare* has hyperonym *cambiare* (to change).

We noticed also other simpler cases: for example for the two senses of the verb *annuire* (to nod) and (to agree), the comparison evidenced that in SIMPLE only the first sense so far was encoded and in IWN only the second one. In similar cases the lack of completeness can be easily mended with a reciprocal advantage. The linking is in fact also a way to spot missing senses, which is particularly useful for IWN, which was built mainly working by taxonomies or semantic classes (i.e. vertically), thus making it difficult sometimes to achieve coverage of all the senses of a word (horizontally).

Finally other problematic cases came in the foreground for these main reasons: duplicated entries in IWN; too fine-grained sense distinctions which made it difficult or impossible to operate the link; synsets containing verbs with a different argument structure (such as {*attaccare, contagiare*} (to infect)). When these kinds of problems are encountered the IWN senses should be revised, corrected in case of duplication, and made easier to be consulted, when necessary, by reducing the sense distinctions.

4. Final remarks

To conclude this report on our experiment, done in the prevision of a future complete mapping between IWN and SIMPLE/CLIPS, first of all we would say that this comparison has been very useful and enlightening about the many and complex problems which we have to deal with when creating lexical resources able to represent in an explicit way (i.e. usable in NLP) semantic information. At our advice it has been of great interest because, forcing us to look at the same word senses / concepts codified according to two different theoretical models of the lexicon, made us more aware of many types of problems. In IWN, a word meaning is analysed in depth and in great detail, thus causing sometimes a too fine-grained sense differentiation and an undesirable proliferation of synsets. SIMPLE concentrates on fundamental senses of a lexical unit, as inferred from the effective usage in the language. This choice has been guided by practical considerations, given that in an applicative context, too fine-grained distinctions risk creating noise, thus compromising the performance of the lexicon e.g. in semantic disambiguation tasks. IWN could benefit by SIMPLE, revising some synsets which result scarcely informative, thus making the resource of more immediate and fluent use; SIMPLE on the other hand could gain focusing the attention on word senses at the moment missing, taking the opportunity to increment the resource itself and to tune the information.

The most interesting result is that, being even more conviced of the usefulness of this linking, a first automatic link should be done starting from the taxonomies belonging to the first order entities. These in fact have nearly always the same ontological patterns; furthermore concrete nouns are less subjected to different interpretations by the lexicographers, being simpler and clearer compared with abstract concepts.

As we showed, many advantages can derive to IWN from the more precise and richer coding of concrete nouns realized in SIMPLE on the basis of qualia structures. On the other hand, the point of view adopted in IWN, based on the centrality of the synset, can be useful to mend a few cases of lack of coherence in the coding of SIMPLE verbs and for the coding of new SemUs.

Obviously the automatic linking, that we plan to realize taking as reference point the isa relations combined with the ontological features, will need a further phase of quite expensive and time-comsuming manual cheking. We know that the automatic linking will be even more difficult for all the second order entities, given the different hyperonyms often assigned in the coding. We think however that this operation will be surely useful to correct the IWN over-differentiation of senses, and that the linking methodology adopted could be profitably reused for the mapping of similar resources.

5. References

Alonge, A., Calzolari N., Vossen P., Bloksma L., Castellon I., Marti T., Peters W. (1998). The Linguistic Design of the EuroWordNet Database. In: Ide N., Greenstein D., Vossen P. (eds.), *Special Issue on EuroWordNet. Computers and the Humanities*, Vol. 32, Nos. 2-3 1998, 91-115.

Calzolari, N., Corazzari, O., Zampolli, A. (2001). "Lexical-Semantic Tagging of an Italian Corpus ". In A. Gelbukh (ed.), *CICLing 2001 Second International Conference on Intelligent text processing and Computational Linguistics Proceedings*, Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg New York.

Corazzari O., Alonge A., Bertagna F., Calzolari N., Roventini A., (2001). "ItalWordNet: extending and exploiting an existing resource for computational tasks". In *NAACL 2001 Workshop, WrodNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, 3-4 June.

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

Lyons, J. (1977). *Semantics.* London, Cambridge University Press.

Miller, G., Beckwith R., Fellbaum C., Gross D., Miller K.J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, Vol.3, No.4, 235-244.

Pustejovsky, J. (1995). *The Generative Lexicon*. The MIT Press, Cambridge, MA.

Rodriguez H., Climent S., Vossen P., Bloksma L., Roventini A., Bertagna F., Alonge A., Peters W. (1998). The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In: Ide N., Greenstein D., Vossen P. (eds.), *Special Issue on EuroWordNet. Computers and the Humanities*, Vol. 32, 2-3, 117-152.

Roventini A., Alonge A., Bertagna F., Calzolari N., Marinelli R., Magnini B., Speranza M. (2002). "ItalWordNet: a Large Semantic Database for the Automatic Treatment of the Italian Language". In: *Proceedings of the First Global WordNet Conference, Central Institute of Indian Languages*, Mysore, India, pp.1-11.

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

Ruimy N., Gola E., Monachini M. (2001). "Lexicography Informs Lexical Semantics: the SIMPLE Experience". In Bouillon P., Busa F. (eds), *The Language of Word Meaning, Studies in Natural Language Processing*, Cambridge University Press, pp. 350-362.

Ruimy N., Monachini M., Distante R., Guazzini E., Molino S., Ulivieri M., Calzolari N., Zampolli A. (2002). "CLIPS, A Multi-level Italian Computational Lexicon: A Glimpse to Data". In *LREC Proceedings 2002*, Las Palmas, Spain.

Ruimy N., Monachini M., Gola E., Calzolari N., Del Fiorentino M.C., Ulivieri M., (forthcoming). "A Computational Semantic Lexicon of Italian: SIMPLE". In *Linguistica Computazionale*, Giardini Editori, Pisa.

Sanfilippo, A., Calzolari N., Ananiadou S., Gaizauskas R., Saint-Dizier P., Vossen P. (eds.) (1999). Preliminary Recommendations on Lexical Semantic Encoding. EAGLES LE3-4244 Final Report.

Vossen, P. (ed.) (1999). EuroWordNet General Document, http://www.hum.uva.nl/~ewn.