Linking Verbal Entries of Different Lexical Resources

Adriana Roventini
Istituto di Linguistica Computazionale Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1- 56124 Pisa, Italy
adriana.roventini@ilc.cnr.it

Abstract

In the field of Computational Linguistics, many lexical resources have been developed which aim at encoding complex lexical semantic information according to different linguistic models (WordNet, Frame Semantics, Generative Lexicon, etc.). However, these resources are often not easily accessible nor available in their entirety. Yet, from the point of view of the continuous growth of technology (Semantic Web), their visibility, availability and integration are becoming of utmost importance. ItalWordNet and PAROLE/SIMPLE/CLIPS are two resources which, tackling lexical semantics from different perspectives and being at least partially complementary, can profit from linking each other. In this paper we address the issue of the linking of these resources focusing on the most problematic part of the lexicon: the second order entities. In particular, after a brief description of the two resources, their different approaches to the verb semantics are described; an accurate comparison of a set of verbal entries belonging to Speech Act semantic class is carried out aiming at evaluate the possibilities and the advantages of a semiautomatic link.

1. Introduction

ItalWordNet (IWN) and PAROLE/SIMPLE/CLIPS (PSC) are two large Italian lexical resources built, in recent years, at the Institute of Computational Linguistics (ILC) in the framework of various international projects (and then enlarged and improved in the national projects *Sistema Integrato per il Trattamento Automatico della Lingua* SI-TAL[1] and *Corpora e Lessici dell'Italiano Parlato e Scritto* CLIPS[2] respectively), according to different lexical semantic models: EuroWordNet (EWN)[3] and the Generative Lexicon respectively. More information about these lexicons, can be found in Roventini et. al. (2003) and Ruimy et al. (2003).

The possibility of using IWN and PSC together, taking advantage of the best features of both underlying models, is the main goal of the linkage we are carrying out. A first survey (Roventini et al. 2002) evidenced the advantages and problems arising from an actual linkage of these resources with regard to concrete entities mostly. In a further step, an exhaustive comparison of the ontologies (Ruimy & Roventini 2005), according to which each resource is structured, allowed to deem a semiautomatic linkage feasible on the whole, even if more problems appeared related to the second order entities, which turned out to be often not easily linkable and need a deeper analysis. To complete this investigation, also in view of the actual realization of this linking in the framework of a project which is now being started at ILC, we carried out a further research on a homogeneous group of verbs (Roventini & Ruimy 2006) with the twofold aim of verifying the envisaged methodology and detecting some other possible problems arising from two different models of lexicon structuring. The work here described continues this research on second order entities testing the mappability of the ontological information to automate the linking.

Manifold advantages are expected from this linking operation. IWN could benefit by the argument structure information encoded in PSC, thus gaining a rich syntactic and semantic subcategorisation and by the extensive domain coding and qualia relations. On the other hand, PSC could take advantage of the extensively encoded synonymy and taxonomy relations of IWN. Furthermore, another advantage for PSC could be the possibility of being related to WordNet through the IWN mapping, thus achieving a multilingual dimension. Finally, both lexicons would gain in coherence and consistency. This linking process can in fact be considered as a sort of reciprocal evaluation of the two resources, and this is particularly important in a field, where inconsistencies, due to lexicographers subjective choices, are hardly avoidable despite the availability of common criteria for coding lexicons.

In the following, the main structural features and the verb semantic coding in both lexicons are briefly described. Then the analysis carried out on the Speech Act verbs in both resources is illustrated together with the preliminary results and the foreseen future work.

2. Outstanding features of both resources

There are a few important differences between these lexicons:

- they are structured in terms of ontologies of a different type – even though partially mappable: PSC has a multidimensional semantic type system organised in a hierarchy, while IWN has a set of rather flat top semantic features;

[1] 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.

[2] http://www.ilc.cnr.it/clips/CLIPS_ENGLISH.htm

[3] 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.

- the basic unit to which all the information is related in PSC is the Semantic Unit (*SemU*), while in IWN it is the *Synset*[4]*;*
- PSC is a lexicon strongly structured by means of templates which provide the semantics of the types ensuring a basic coherence of coding;
- IWN model is noticeable rich in semantic relations, but its little structured representation of information favors an uneven distribution of it.

Sketching out, a different philosophy permeates these lexicons according to the different theoretical models they refer to: WordNet (Miller et al. 1990) and the Generative Lexicon (Pustejovsky, 1995). In IWN the richness of sense distinction and the variety of semantic relations holding among the synsets are put in the foreground while PSC's outstanding features are the connection between syntax and semantics and a rigorous method of lexicon structuring.

As it will be evident in the following, IWN inherited from the WordNet model, in particular for verbal entries, the proliferation of slightly different senses associated with a lexical item. In PSC, by contrast, more generic and less numerous senses are encoded for a lemma.

3. Semantic representation of verbs in IWN

Taking as models both Princeton WordNet and Cruse's approach (Cruse 1986) to meaning representation, a relational view of the lexicon characterizes IWN lexical model (Alonge et al 1998) according to which all the semantic aspects regarding the lexical level are reflected in the paradigmatic and syntagmatic relations existing between any two words in a language. Therefore, the meaning of a word is described both in terms of other words displaying a similar meaning in a specific context (synonymous) and by referring to the relations that a word has with the other words in the lexicon, i.e. to its location within a net. Many lexicalization patterns of 'semantic components' are encoded, whenever possible, without drawing a sharp distinction between what is strictly speaking 'semantic' and what could be described as 'pragmatic meaning'.

This can be seen in particular in the verb coding, where the INVOLVED relation is used to encode data on arguments or adjuncts lexicalized within the meaning of a verb. This relation links a verb and a first order noun whose meaning is connected with the verb itself[5]. Furthermore specific subtypes of this relation (AGENT, PATIENT, INSTRUMENT, LOCATION) make this relation particularly useful.

As far as the ontology structure is concerned, in IWN there is a hierarchy of 60 language-independent top concepts, reflecting fundamental semantic distinctions, built within EWN and partially modified in SI-TAL project, to account for adjectives. The verbs, as entities belonging to the second order (Lyons 1977), are organized in two classification schemes, which represent the first division below this order: *Situation Type* and *Situation Component* (cf. Table 1). The *Situation Type* is connected with the event-structure or *Aktionsart* (lexical aspect) of a situation distinguishing two aspects: Static and Dynamic (which in its turn has as subtypes BoundedEvent and UnboundedEvent). The *Situation Component* lists 22 salient semantic components that characterize situations. In IWN, each verb synset is marked by one well defined and precise *Situation type* to which many different combinations of *Situation Component* concepts are associated.

2nd ORDER ENTITY
SITUATION COMPONENT
Cause
Communication
Condition
Existence
Experience
Location
Manner
Mental
Modal
Physical
Material
Physiological
Possession
Purpose
Quantity
Social
Time
Intensity
Property
Attribute
Functional
Relation
SITUATION TYPE
Dynamic
BoundedEvent
UnboundedEvent
Static

Table 1 IWN top concepts for events

4. Semantic representation of verbs in PSC

In the PSC lexicon, the semantic content of a verb is expressed by its membership in a semantic type (cf. Table 2) to which a rich bundle of semantic features and relations is associated. Among these there are the 60 relations of the Extended Qualia structure, an enlarged version of the GL representational tool that enables to describe the componential aspect of a word meaning as well as its relationships to other lexical items.

The semantic description of verbs also encompasses contextual information, formulated in terms of a semantic predicate and its arguments with their thematic roles and semantic typing. Syntactic and semantic information

[4] Each synset is constituted by various synonyms gathered according to the weak definition of synonymy adopted in WordNet and consequently in IWN, stating that "two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value".

5 The relation *Role* is used for the opposite link, from concrete nouns to verbs (or nouns referring to states, processes or events).

Event
- Phenomenon
  - Weather verb
  - Disease
  - Stimulus
- Aspectual
  - Cause aspectual
- State
  - Exist
  - Relational state
    - Identificational state
    - Constitutive state
    - Stative location
    - Stative possession
- Act
  - Non_relational_act
  - Relational_act
    - Cooperative_activity
    - Purpose_act
  - Move
    - Cause_motion
  - Cause_act
  - Speech_act
    - Cooperative_ speech_act
    - Reporting_event
    - Commissive_speech_act
    - Directive_speech_act
    - Expressive_speech_act
    - Declarative_speech_act
- Psychological_event
  - Cognitive_event
    - Judgement
  - Experience_event
  - Cause_experience_event
  - Perception
  - Modal_event
- Change
  - Relational_change
    - Constitutive_change
    - Change_of_state
    - Change_of_value
  - Change_of_possession
    - Transaction
  - Change_of_location
  - Natural_transition
  - Acquire_knowledge
- Cause_change
  - Cause_Relational_change
    - Cause_Constitutive_change
    - Cause_Change_of_state
    - Cause_Change_of_value
  - Cause_Change_location
  - Cause_Natural_transition
  - Creation
    - Physical_creation
    - Mental_creation
    - Symbolic_creation
    - Copy_creation
  - Give_knowledge

Table 2 SIMPLE-CLIPS semantic types for events

concerning a verb is linked through the projection of the predicate-argument structure onto its syntactic realization(s). A basilar element in PSC semantic encoding is the *template*, i.e. a schematic structure which allows to constrain a semantic type to a structured cluster of information considered crucial to its definition and eases the lexicographer's task, thus enhancing the consistency and structuring the linguistic information encoded. The PSC ontology, part of which can be seen in Table 2, was conceived and set up in the EU *LE-SIMPLE* project[6]. It is more structured and detailed compared to the IWN one, and for this reason it is taken as a point of reference for the semiautomatic linking.

5. The Speech Act verb class

The semiautomatic link we are planning avails itself of both the 'is-a' or hyperonymy relation and the ontological concepts or top concepts (EWN) or semantic types (PSC), taking as reference point the SIMPLE ontology, which allows the extraction of more coherent ad homogeneous sets of verbs.

The first study (Roventini & Ruimy 2006) on verb merging was carried out on more than one hundred verbal entries both causative and inchoative belonging to the 'feeling' semantic field and, given the encouraging results obtained, we decided to widen the research and verify this methodology taking the Speech Act verbs as a further testing bench.

According to this procedure, all the PSC verbal SemUs, belonging to the Speech_Act hierarchy of semantic types, were selected to be compared with the corresponding IWN synsets.

In PSC Speech Act verbs are distributed over six different templates (cf. Table 2). In IWN the basic top concept according to which Speech Act verbs are classified is Communication. This top concept then combines with other *Situation Components* and *Situation Type* top concepts as shown in Table 3.

On the basis of their ontological classification the Speech_Act verbs were extracted from PSC database to be compared with the corresponding IWN synsets and analyzed. In all 188 SemUs were extracted corresponding to 155 synsets and to 300 variants[7]. Afterwards a manual check has been performed and 131 synsets out of 155 (about the 84%) turn out to be good candidates for the matching.

Top concepts	**Synsets**
Agentive Communication Dynamic	97
Agentive BoundedEvent Communication Purpose	13
Agentive Communication Purpose	9
Agentive Communication UnBoundedEvent	8
Agentive Communication Dynamic Social	2
Agentive BoundedEvent Comm. Mental Purpose	2
Total	131

Table 3 Semantic concepts in IWN for speech verbs

As regards the correspondence existing among the IWN top concepts and PSC semantic types, it has been observed

[6] http://www.ub.es/gilcub/SIMPLE/simple.html

[7] Each synset is constituted by 1 to n word sense, called variant in the EWN terminilogy.

that, in general, Agentive Communication UnBoundedEvent corresponds to Cooperative_Speech_Act and Agentive BoundedEvent Communication Purpose to Directive_Speech_Act, but, for the most part, the IWN verbal synsets belonging to the Speech Act class are marked out as Agentive Communication Dynamic.

The analysis carried out on this class confirms that in IWN the ontological classification is less precise and, sometimes, incomplete or under specified as regards the S*ituation Type*. In fact if we consider the data in Table 3 we notice that the specification, BoundedEvent or UnboundedEvent, is present in 23 synsets out of 131. This fact can be attributed to little precision in the choice of hyperonyms when coding verbal entries, which entailed a consequent lack of precision at the ontological level.

5.1. Analyzing sense distinction in IWN and PSC

One of the major problems which emerged when we started considering the linking of the two resources, is the different granularity of sense distinction, especially when merging verb classes. IWN, on one hand, tends to over detail the senses of a lexical item and to combine as many synonyms as possible within a synset; PSC, on the other hand, only accounts for fundamental meaning distinctions. This imbalance came out both when analyzing the "feeling" verb subset and in this further investigation on Speech Act verb class. Nevertheless we are convinced that a good harmonization of the resources can overcome this difference, turning it into an advantage.

As regards this problem we analyzed some of the selected entries. In Tables 4 and 5 below the coding of *parlare* (to speak, to talk) in both resources is shown.

In Table 4 hyperonyms and semantic types encoded in PSC for the lemma *parlare* are shown: in PSC this verb has 3 senses and no synonym is indicated. This little number of senses is probably due to the strict connection between syntax and semantics, in any case here we find one SemU for each different argument structure, while more subtle meaning distinctions are not taken into consideration.

In IWN the lexical item *parlare* appears in 11 synsets where it is associated to 11 synonyms (out of which 2 are verbal multiwords). Table 5 evidences both the different granularity of sense distinction and the under specification of the ontological typing. In fact 7 out of 8 Speech Act synsets are marked out as Agentive Communication Dynamic, which is the more general combination that characterizes Speech Act verbs in IWN.

If we consider the PSC senses compared to the IWN ones we notice that the first PSC sense, USem4876, matches one sense of IWN, synset 33942, trough the hyperonym *sapere* (to know). On the basis of both semantic type and hyperonym relation the second PSC sense, USem67407, matches the sixth IWN sense, synset 33938. The third PSC sense instead, USemD439, does not match the corresponding IWN sense, i.e. synset 33933. This happens because the eleventh sense of *parlare* in IWN, synset 33943, has the same hyperonym *dire* which is misleading and makes the disambiguation impossible.

The other IWN senses are due in some cases to very subtle meaning distinctions not accounted for in PSC. For example synset 33935, which indicates the human ability of using articulate language, or synset 33944 which is a figurative sens, or synset 33939 which means *parlare in pubblico, tenere un discorso* (give a speech, address). The synset 33937 could map, through the hyperonym the Usem62401, but in PSC pronominal forms are not retrievable as Usem. The remaining three synsets, according to the IWN coding, do not belong to Speech Act. Summing up we find three equivalent senses, but only two can match automatically. This case is reported as an example of high degree of imbalance between the two resources, in many other cases we found a fairly good correspondence in sense distinction and reciprocal enhancement.

Hyperonym (isa relation)	Semantic Type	SemU identifier
Sapere (to know)	Cognitive_event	USem4876
Comunicare (to comunicate)	Cooperative_speech_act	USem67407
Dire (to say)	Speech_Act	USemD439

Table 4 Hyperonyms and semantic types in PSC for *parlare*

Hyperonym (isa relation)	Ontology Concept	Synset dentifier
Comunicare 2 (to communicate	Agentive Comm.Dynamic	33933
Potere 1 (to can)	Agentive Comm.Dynamic	33935
Esprimere 2 (to express)	Agentive Comm.Dynamic	33944
Rivelare 3 (to reveal)	Cause	33936
*Pronunciars*i 1(to judge)	Agentive Comm.Dynamic	33937
Comunicare 1(to communicate)	AgentiveComm.UnboundedEvent	33938
Parlare 1(to talk, speak)	Agentive Comm.Dynamic	33939
Palesare 1(to make_known)	Cause	33940
Progettare 1(to plan)	AgentiveExist.MentalPurpose	33941
Sapere 1 (to know)	Agentive Comm.Dynamic	33942
Dire 1 (to say)	Agentive Comm.Dynamic	33943

Table 5 Hyperonyms and ontology concepts in IWN for *parlare*

5.2. A few cases of reciprocal enhancement

In a joint consultation of these lexicons much more lexical information will be available as proved by the following examples.

Let consider for example the lexical item *calunniare* (to calumniate).

In IWN it has one sense represented by the multi variants synset 32365 {*calunniare, denigrare, detrarre diffamare, infamare*, *vituperare*} (to calumniate, defame, denigrate, slander, smirch, asperse, smear, sully, besmirch, charge_falsely..)[8], it is linked to *dire* (talk, utter, speak, mouth, verbalize) by a hyperonym relation and shows the ontological classification: Agentive Communication Dynamic.

In PSC the corresponding SemUs are encoded in the following way: *calunniare* is an Expressive_speech_act linked by an is-a relation to *dire; detrarre*, *diffamare*, *infamare*, are all hyponyms of *calunniare* and belong to semantic type Expressive_speech_act; *denigrare* also belongs to Expressive_speech_act, but no is-a relation is indicated, *vituperare* is encoded as Expressive_speech_act and linked by an is-a relation to *offendere* (to offend). Comparing all variants of the IWN synset 32365 to the corresponding SemUs we find a precise correspondence with Usem60794*calunniare.* As regards UsemD65971*detrarre,* Usem60878*diffamare* and Usem66192*infamare* they appear hyponyms instead of synonyms of *calunniare* but their semantic type makes it possible to automatically match them. Also the UsemTH295*denigrare* matches the IWN synset through its ontological classification, proving how a correct choice of the semantic type is of the utmost importance. As regards the last variant, *vituperare*, it matches UsemD65724*vituperare* thanks to the ontological classification, since the is-a relation in PSC points to *offendere* (to offend) instead of *calunniare*. This discrepancy is cleared up by the comparison, which evidences that *vituperare* has two meanings. In fact, in IWN, *vituperare* is also member of the synset 32313 {*offendere, ingiuriare, insultare, oltraggiare, vilipendere vituperare* } (offend, insult, affront, hurt, wound, injure, spite) marked out as Dynamic Experience Mental Stimulating. The PSC UsemD65724 *vituperare* combines two different senses: as Speech_Act it should be linked by an is_a relation to *calunniare*, while the hyperonym *offendere* should require the semantic type Cause_Experience_Event which corresponds to the IWN top concepts combination Dynamic Experience Mental Stimulating.

Another example showing the reciprocal enhancement deriving from a linking process is constituted by the synset 35281{*beffare, sbeffare, corbellare, berteggiare, dileggiare, deridere, irridere, schernire, sbeffeggiare, sbertucciare, sfottere, prendersi_gioco*}(mock, jeer, scoff, flout, barrack, gibe). Compared to the corresponding PSC Usems, on the one hand it provides information on both many synonymy relations and word senses not present in PSC i.e. *sbertucciare, berteggiare, sfottere, corbellare, sbeffeggiare*; on the other hand, it appears not well formed. In fact it includes a multiword expression, *prendersi_gioco*, incompatible with the other variants in the synset as for argument structure. Many other examples could be reported, but the most frequent types of reciprocal enhancement are the ones just described: a better information on synonymy and a greater richness of senses is provided by IWN, more detailed ontological classification and rigorous attention to the relations existing between syntax and semantics is provided by PSC.

6. Final remarks

In this paper we have described a detailed analysis aimed at investigating the possibility of semi-automatically linking the two largest and richest Italian lexical resources, IWN and PSC, as far as second order entities are concerned.

The methodology adopted, which is grounded on the mapping of both hyperonymy relations and ontological classification, was firstly experimented on a set of causative and inchoative verbs of "feeling". The results of this previous test appeared promising enough to encourage us to carry on our linking project. To complete this investigation on second order entities, in view of the actual realization of this linking, we carried out a further test on the verb class of Speech Act with the aim of verifying the envisaged methodology and detecting some other possible problems.

We are now even more convinced of the viability of such a linking, given the results we obtained for this class. In fact, while for the set of the "feeling" verbs about the 50% of the IWN synsets were found linkable with correspondent PSC entries, Speech Act verbs exceeded our expectations since about 84% of them are good candidate for a correct linkage. On the basis of this new encouraging result, we intend to complete the comparison of second order entities in a semiautomatic way.

Given the smaller number of verbal entries and the greater homogeneity of coding guaranteed by the PSC templates, we will proceed in the comparison extracting, one semantic type at time, the verbal SemUs from PSC and matching them to the corresponding synsets in IWN. Once completed the automatic extraction of the matched couples, the candidate joint entries will be checked for adjustments and harmonization. By means of this procedure we expect to be able to link the most part of verbal PSC SemUs with a corresponding IWN synset, and to circumscribe in this way an intersection set of verbs showing the most valuable features of both resources**.** Much more information will be available, such as the argument structure, a more essential sense distinction usable by automatic systems of natural language processing, a more precise ontological description, the possibility of exploiting the methodical coding of synonymy and the link to WordNet.

[8] In the round brackets is entirely reported the WN 1.5 entry, to which the Italian synset is linked.

7. References

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

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

Lenci, A., Bel, N., Busa, F., Calzolai, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Petres, W., Ruimy, N., Villegas, M., Zampolli, A. (2000). "SIMPLE: A General Framework for the Development of Multilingual Lexicons", «International Journal of Lexicography», XIII(2000) 4: 249-263.

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

Miller, G., Beckwith, R., Fellbaum C., Gross D., Miller K.J., *Introduction to WordNet: An On-line Lexical Database*, «International Journal of Lexicography», III (1990), 4, 235-244.

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

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: «Linguistica Computazionale», vol. XVI-XVII pp.745-791, Giardini, Pisa.

Roventini, A., Ulivieri, M., Calzolari, N. (2002). Integrating two semantic lexicons, SIMPLE and ItalWordNet: what can we gain?. In: LREC, Proceedings of the Third International Conference of Language Resources and Evaluation, Vol. V, pp. 1473-1477.

Roventini, A., Ruimy N. (2006). Linking and harmonizing different lexical resources: a comparison of verbal entries in ItalWordNet and PAROLE-SIMPLE-CLIPS. In Proceedings of the Third International WordNet Conference, Jeju Island, Korea

Ruimy, N., Battista, M., Corazzari, O., Gola, E., Spanu, A., (1998). Italian Lexicon Documentation, LE-PAROLE, Final Review Documentation, LE-4017 Parole Project, April.

Ruimy N., Monachini M., Gola E., Calzolari N., Del Fiorentino M.C., Ulivieri M., Rossi S. (2003). A computational semantic lexicon of Italian: *SIMPLE* . In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa. *Linguistica Computazionale*, Special Issue, XVIII-XIX. Pisa-Roma, IEPI. Tomo II, 821-864.

Ruimy N., Roventini A. (2005). Towards the Linking of two Electronic Lexical Databases of Italian. In Z. Vetulani (ed.), L&T'05, April 21-23, Poznan, Poland. Wydawnictwo Poznanskie Sp. z o.o. 230-234.

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

.