ItalWordNet: a Large Semantic Database for Italian

Adriana Roventini*, Antonietta Alonge**, Nicoletta Calzolari***

Istituto di Linguistica Computazionale, CNR Area della Ricerca di Pisa Via Alfieri 1, Loc. S. Cataldo Ghezzano 56010 (PI) – ITALY * adriana@ilc.pi.cnr.it, ** antoalonge@libero.it, *** glottolo@ilc.pi.cnr.it

Bernardo Magnini

Istituto per la Ricerca Scientifica e Tecnologica I-38050, Povo, Trento magnini@irst.itc.it

Francesca Bertagna

Consorzio Pisa Ricerche Via S. Maria 40 Pisa 56100 - ITALY F.Bertagna@ilc.pi.cnr.it

Abstract

The focus of this paper is on the work we are carrying out to develop a large semantic database within an Italian national project, SI-TAL, aiming at realizing a set of integrated (compatible) resources and tools for the automatic processing of the Italian language. Within SI-TAL, ItalWordNet is the reference lexical resource which will contain information related to about 130,000 word senses grouped into synsets. This lexical database is not being created *ex novo*, but extending and revising the Italian lexical wordnet built in the framework of the EuroWordNet project. In this paper we firstly describe how the lexical coverage of our wordnet is being extended by adding adjectives, adverbs and proper nouns, plus a terminological subset belonging to the economic and financial domain. The relevant changes involved by these extensions both in the linguistic model and in the data structure are then illustrated. In particular we discuss i) the new semantic relations identified to encode information on adjectives and adverbs ii) the new architecture including the terminological subset.

1. Introduction

Since the Princeton WordNet database, a semantic network in which the meanings of words are represented in terms of their conceptual-semantic and lexical relations to other words (Miller et al., 1990), has become available it has been the tool of choice for researchers aiming at building Natural Language Processing (NLP) systems of various kinds, mainly because that resource contains information which is necessary for a fundamental task of various applications, i.e. for Word Sense Disambiguation (WSD). However, WordNet was not designed to meet the requirements of NLP, so it has become clear that it lacks some information which could be very useful for a variety of applications (cf. Gonzalo et al., 1998).

The main goal of the EuroWordNet (EWN) project¹ was thus to develop a (multilingual) lexical resource,

retaining the basic underlying design of WordNet (in particular, of the database version WordNet 1.5, hereafter WN1.5) while at the same time trying to *improve* it in order to answer the needs of research in the computational field. Thus, a fundamental change made in EWN was that the set of lexical relations to be encoded between word meanings was extended or modified in various ways with respect to the set defined in WN 1.5.

Within EWN semantic information was encoded, for about 50,000 word senses (nouns and verbs) in each of the languages dealt with, in the form of lexical semantic relations between synsets (i.e. synonym sets). A rich framework of relations was designed which were considered useful for computational applications. In any case, synonymy and hyponymy were extensively encoded in the wordnets produced, while, due to time limits, the more 'sophisticated' relations were encoded for selected sets of words in each wordnet.

We are now working at extending the wordnet produced for Italian, inserting adjectives and adverbs, but also nouns and verbs which had not been taken into consideration yet in EWN. This is being done in the context of a National Project which aims at building various integrated language resources for the automatic

¹ EWN was a project in the EC Language Engineering (LE4003) programme. In a first phase, the partners involved were the University of Amsterdam (coordinator); the Istituto di Linguistica Computazionale, CNR, Pisa; the Fundacíon Universidad Empresa (a cooperation of UNED, Madrid, Politecnica de Catalunya, Barcelona, and the University of Barcelona); the University of Sheffield; and Novell Linguistic Development (Antwerp), changed to Lernout & Hauspie during

the project. In a further phase, the database was extended with German, French, Estonian and Czech.

treatment of the Italian written and spoken language². Among such resources, a very large semantic network is being developed (ItalWordNet - IWN) which will contain information related to 80,000 synsets (about 130,000 word senses), linked to WN1.5. Although we are basically using the EWN model of lexical-semantic relations to build it, we are identifying additional relations, mainly to be used to encode data on adjectives (which were encoded in EWN only as targets of relations from nouns and verbs). Moreover, we are adding a terminological subset, related to economy, in such a way that it will possible to access either only the generic lexicon in the database or the specialized set, or also both the subsets at the same time. In this paper, we describe the overall architecture of the IWN database, discussing in particular the new relations encoded and the characteristics of being the terminological subset.

2. The overall architecture of the IWN database

The IWN database is constituted by:

- a generic wordnet, built by extending the network developed within EWN, which will contain, at the end, about 130,000 word senses corresponding to about 80,000 synsets;
- a (generic) Interlingual-Index (ILI) which is an unstructured version of WN1.5, i.e. it contains all the synsets found in WN1.5 but not the relations among them. This module was used in EWN to link wordnets of different languages. In IWN we also link the synsets encoded in the generic wordnet to such an interlingua, to make the resource usable in multilingual applications;
- iii) a terminological wordnet, containing synsets found in the economic-financial domain. This will be linked to the generic wordnet, as will be explained below;
- iv) a terminological ILI, containing synsets related to the economic-financial domain, partly extracted from $WN1.6^3$;
- v) the Top Ontology (TO), a hierarchy of languageindependent concepts, reflecting fundamental semantic distinctions, built within EWN and partially modified in IWN to account for adjectives (which were not dealt with in EWN). The TO consists of language-independent features which may (or may not) be lexicalized in various ways, or according to different patterns, in different languages (Rodriguez et al., 1998). Via the ILIs, all the concepts in the generic and specific wordnets are directly or indirectly linked to the TO;
- vi) the Domain Ontology (DO), containing a set of domain labels. In EWN this module was only partially developed and used to code information on computer terminology, whereas in IWN a complete set of labels is being developed. Via the ILIs, all the

concepts in the generic and specific wordnets are directly or indirectly linked to the DO.

The overall architecture of IWN is shown in Figure 1 below.



Figure 1: The overall architecture of IWN

3. The IWN Linguistic Model

3.1 Lexical Coverage of IWN

To extend the lexical coverage of the resource developed within EWN, that is being used as the first nucleus of IWN, various resources were analysed, and in particular: the Italian Machine Dictionary or DMI⁴, the PAROLE⁵ lexicon, the lemmatised corpus selected for the semantic annotation within SI-TAL⁶, the Italian Reference Corpus⁷. From this survey, the main source of new lemmas and relative senses turned out to be the DMI. In fact, by means of a comparison of the already developed wordnet with the DMI we were able to identify about 23,000 new senses for the nouns and 6,000 new senses for the verbs.

The comparison with the PAROLE lexicon and the TAL corpus showed a nearly complete coverage as far as verbs and nouns are concerned. In Table 1 below are reported the results of the comparison with PAROLE.

² The SI-TAL ('Integrated System for the Automatic Treatment of Language') National Project.

³ While we use version 1.5 of WN for the generic ILI, because this was the version used to build the wordnet developed within EWN, constituting the nucleus of the IWN generic wordnet, the specific wordnet is being linked to WN1.6 because in this version many economy terms are encoded.

⁴ This dictionary was realized in the seventies and contains about 106,000 lemmas, more than a million of inflected word-forms and 187,000 definitions for the three main parts of speech. Firstly stored on tapes it was then changed into a relational database on main frame and finally transferred on pc (cf. Calzolari, Ceccotti & Roventini, 1983).

⁵ The PAROLE LE 40-17 (Preparatory Action for Linguistic Resources Organisation for Language Engineering) project was funded by the European Community to create a first set of compatible corpora and lexicons for eleven different European languages.

⁶ This corpus is divided into two different subsets: the "balanced corpus" constituted by various types of texts and the "financial corpus" which only contains texts belonging to the selected economic – financial domain.

⁷ The Italian Reference Corpus, developed at the Institute for Computational Linguistics in Pisa utilizing texts of various types, contains about 12,000,000 of word-forms (cf. Bindi et al., 1989).

PAROLE entries	EWN entries	Intersection Subset	EWN only	PAROLE only
Nouns	Nouns	Nouns	Nouns	Nouns
13257	24416	11377 (85.82%)	13039	1880
Verbs	Verbs	Verbs	Verbs	Verbs
3090	6578	2868 (92.82%)	3710	222

Table 1: Comparison between EWN and PAROLE

Moreover, within EWN we did not deal with adjectives and encoded only a few proper nouns, while, of course, in order to be able to use our resource for WSD⁸ or other tasks, we had to add them in IWN. So, we firstly identified a set of adjectives to be encoded in our database and chose some adverbs (those derived by adjectives by means of the suffix *-mente*) to be added too. Then, given the high incidence/frequence of proper nouns not only within the TAL corpus but in whatever other corpus/text, we considered as an useful extension the introduction of a few types of proper nouns, in particular those which have common nouns as derivatives. According to this criterion of derivational productivity, we are introducing in IWN a set of istances belonging to both the geografic and human domains.

As far as the adjectives are concerned, we selected from the DMI about 17,000 senses already partially encoded by means of numeric codes which individuate different types of definitions such as the synonymical or the functional type (Calzolari, Ceccotti & Roventini, 1983). This cluster has been the starting point of our analysis for the linguistic model of this class (see the following subsection). Below we show a few examples of the definitions encoding we benefit when acquiring this class and defining for it the new semantic relations. In the examples a) and b) we have two synonymical definitions in which the antonym(s) is (are) also indicated (signalled by C.=); in the examples c) and d) we have two functional definitions in which, the defining formula "atto a", points out the regular semantic relation that the suffix -ivo establishes between a large group of verbs and the derived adjectives.

a) *Balsamico* (balsamic): *odoroso* (fragrant), *salubre* (healing) C = fetido (foetid, stinking), *malsano* (unhealthy)

b) *Benevolo* (benevolent): *benigno* (well-disposed), *affabile* (affable) *C*.= *malevolo* (malevolent)

c) *Disgiuntivo* (disjunctive): *atto a* (suitable for) *disgiungere* (to disjoin)

d) *Elogiativo* (laudatory): *atto a* (suitable for) *elogiare* (to praise)

In addition to the DMI definitions, the contexts of the most frequent adjectives occurring in the TAL corpus have been analysed. This preliminary study of the information contained in both the dictionary definitions and the corpus contexts made it possible to individuate a few Base Concepts for the adjectives, following the same criterion used for nouns and verbs in the EWN framework: i) high number of relations with other dictionary entries, ii) high usage frequency⁹. As the nouns and verbs Base Concepts, the most frequently used adjectives are characterised by generic meaning and high degree of polisemy. Moreover, they show aptitude to combine with various types of nouns. These, together with the adjectives showing a considerable number of hyponyms, will be the Base Concepts of this class.

3.2 Lexical-semantic relations being encoded

The basic notion around which the IWN database is built is the same around which both WN and EWN were built, i.e. that of a *synset*, or set of synonymous words with the same Part-of-Speech (PoS) that can be interchanged in a certain context. Then, the network is mainly based on the hyponymy (or IS-A) relation, but various other relations can be encoded, partly inherited from EWN, to better describe the semantics of words.

Whereas in WN1.5 a rigid distinction is drawn among different PoSs and each PoS forms a separate system of language-internal relations, following EWN in IWN various relations applying between different PoSs can be encoded. Indeed, instead of separating the networks on the basis of their PoSs, traditionally identified by using a mixture of morphological, syntactic and semantic criteria, a distinction is drawn among the semantic orders of the entities to which word meanings refer (cf. Lyons, 1977): I^{st} order entities (referred to by concrete nouns), 2^{nd} order entities (referred to by verbs, adjectives or nouns indicating properties, states, processes or events), and 3^{ra} order entities (referred to by abstract nouns indicating propositions existing independently of time and space). This allows to establish cross-PoS relations between words of the same semantic order referring to similar concepts (e.g., cross-PoS synonymy between arrival and to arrive, etc.), but also other cross-PoS relations which emphasize the language-specific lexicalisation patterns of semantic components. Table 2 provides a list of the main relations defined in EWN which are also encoded in IWN.

⁸ As said above, within SI-TAL IWN is the reference lexical resource and, in this framework, it is used mostly for the corpus semantic annotation. Since in the SI-TAL corpus we find 3.370 adjectival lemmas which are not encoded in EWN, and 43,80% of the nouns found are proper nouns we needed to deal with the issue of properly encoding both adjectives and proper nouns in IWN.

⁹ Within EWN a common set of so-called Base Concepts were defined for nouns and verbs and used as a starting point by all the sites to develop the cores of the wordnets. Base Concepts are meanings that play a major role in the wordnets.

¹⁰ This approach can yield some remarkable advantages with respect to the use of the database both for IR purposes and for other LE applications (cf. Gonzalo et al., 1998).

Relation	Semantic orders linked	Example
NEAR_SYNONYM	1st/1st, 2nd/2nd, 3rd/3rd	tools/instrument; to bear witness/to assure
XPOS_NEAR_SYNONYM	2nd/2nd	arrival/to arrive
ANTONYM	1st/1st, 2nd/2nd, 3rd/3	inside/outside; arrival/departure; low/high
XPOS_ANTONYM	2nd/2nd	arrival/to leave
HAS_HYPERONYM/HAS_HYPONYM	1st/1st, 2nd/2nd, 3rd/3	dog/animal; to arrive/to go
XPOS_HAS_HYPERONYM/HYPONYM	2nd/2nd	arrival/to go
HAS_HOLONYM/HAS_MERONYM	1st/1st	arm/body; hand/finger
CAUSES/IS_CAUSED_BY	2nd/2nd	to kill/to die; to execute/death sentence
HAS_SUBEVENT/IS_SUBEVENT_OF	2nd/2nd	to buy/to pay; to snore/to sleep
INVOLVED/ROLE	2nd/1st or 3 rd and viceversa	to hammer/hammer; to enter/inside
CO_ROLE	1st/1st	guitar player/guitar
BE_IN_STATE/STATE_OF	1st/2nd and viceversa	poor (N)/poor (Adj)
IN_MANNER/IS_MANNER_FOR	2nd/2nd (Adv)	to whisper/in a low voice
HAS_INSTANCE/BELONGS_TO_CLASS	1st/1st (proper noun)	river/Thames
DERIVATION	All (between lexical units)	presidential/president

Table 2: Main IWN relations in	nherited from	EWN
--------------------------------	---------------	-----

Note that some of these relations can be further specified (e.g., the verb *uscire* – to go out – may be linked to the noun *esterno* – outside – by means of an INVOLVED_ TARGET_DIRECTION relation). Moreover, some labels may be added to certain relations to make clear some implications that they may carry: *conjunction* and *disjunction* (for multiple relations of the same kind encoded for a synset); *non-factivity* (to indicate that a causal relation does not necessarily hold); *intention* (added to a cause relation to indicate intention to cause a certain result); *negation* (to explicitly encode the impossibility of a relation occurring); *reversed* (automatically added by the tool to reversals of not conceptually bi-directional relations).¹¹

Within IWN we have further enriched the set of relations which can be encoded, mainly to account for adjectives and adverbs (which were not dealt with in EWN), but also to encode certain relations more properly. For instance, within EWN a whole family of 'causal' relations was identified:

- a) *to kill* RESULTS_IN *to die*
- b) to search FOR_PURPOSE_OF to find
- c) vision ENABLES_TO to see
- d) *heat* IS_MEANS_FOR *to distill*

However, given the granularity of such distinctions and the time limits of the project, an underspecified CAUSE relation, together with labels indicating particular implications, was actually used in EWN to encode all these links. In IWN, we have instead decided to encode (a), (b) and (d) relations above as sub-classes of the underspecified CAUSE relation, due to their prominence, while we shall not use (c) relation, since it only applies between a very limited number of synset pairs.

Much work has then been devoted at identifying a set of relations to be used to encode data on adjectives, since within EWN no analysis had been carried out on this topic. In WN the possibility of encoding hyponymy for adjectives is denied and the basic relation encoded for adjectives is antonymy. Within IWN we have reconsidered the possibility of encoding hyponymy for adjectives, given the important inferences which can be drawn on the basis of this relation. By analysing data coming from machine-readable dictionaries (in particular from the DMI) we find subsets of adjectives which have a *genus* + *differentia* definition, like nouns or verbs. That is, these adjectives seem to be organised into classes sharing a superordinate. Here below some examples are provided:

albino (whitish): *affetto* da albinismo (suffering from albinism) acneico (acned): *affetto* da acne (suffering from acne)

acquoso (watery): *contenente* acqua (containing water) alcalino (alkaline): *contenente* alcali (containing alkalis)

filoso (thready): *pieno* di fili (full of threads) stellato (starry): *pieno* di stelle (full of stars).

We have decided, therefore, to encode hyponymy also for these sets of adjectives. The IS-A taxonomies which can be built are different from those built for nouns or verbs, since they are generally rather flat, consisting almost always of two levels only (an exception is, e.g., the color adjectives taxonomy). However, by encoding hyponymy for these adjectives, we obtain classes for which it will be possible to make various inferences. For instance, it will be possible to infer semantic preferences of certain classes: e.g., all the adjectives occurring in the taxonomy of affetto above will modify nouns referring to animate entities; the *contenente* hyponyms will occur as attributes of concrete nouns; etc. Furthermore, it will also be possible to infer information on syntactic characteristics of adjectives found in the same taxonomy: e.g., the hyponyms of atto (suitable for) are always found in predicative position (and do not accept any complements); the hyponyms of privo (lacking) may occur both in attributive and in predicative position (and may take certain prepositional complements), etc..

As in WN, also in IWN the antonymy relation remains an important relation to describe the semantics of various adjectives. Following theoretical work (Lyons, 1977;

¹¹ We are not going into the details of all the EWN relations here. For a complete discussion of them see Alonge et al. (1998).

Cruse, 1986), we have further distinguished between COMPLEMENTARY ANTONYMY and GRADABLE ANTONYMY¹². The former relation links synsets referring to opposing properties/concepts: when one holds the other is excluded (alive/dead). The latter relation is used for those antonym pairs which refer to gradable properties (long/short). This information can be useful for computational applications since word pairs presenting one of the two kinds of opposition may occur in different contexts (cf. Cruse, 1986). However, in case it is not clear if two opposing synsets should be linked by COMPLEMENTARY_ANTONYMY or GRADABLE_ ANTONYMY, we can still use the underspecified ANTONYMY relation.

In WN a relation between adjectives and nouns is encoded for relational adjectives which point to a noun to which they 'pertain': *atomic/atom*, *industrial/industry*, etc. This relation will also be encoded in IWN, by using the label PERTAINS_TO.

Another relation 'inherited' from WN can be useful to distinguish both adjective senses and their semantic preferences:¹³

alto ₁ (tall)	IS_A_VALUE_OF	statura (stature)
alto ₂ (high)	IS A VALUE OF	altezza (height).

A new relation, not present either in WN or in EWN, will be encoded for a class of adjectives indicating the possibility of some events occurring:

giudicabile (= che può essere giudicato) (triable) LIABLE_TO giudicare (to judge)

Since adjectives are 2^{nd} order entities, we may encode for them relations used for the other 2^{nd} order entities. In particular, we encode the 'INVOLVED' and 'CAUSE' relations. The INVOLVED relation links a 2^{nd} order entity with a 1^{st} (or 3^{rd} order) entity referring to a concept *incorporated* within the meaning of the 2^{nd} order entity (cf. Table 2). Examples for adjectives are given in the following:

stellato (= pieno di stelle) (starry) HAS_HYPERONYM pieno (full) INVOLVED stella (star) squamato (= ricoperto di squame) (scaly) HAS_HYPERONYM ricoperto (covered) INVOLVED squama(scale). The CAUSE relation and its sub-relations link 2^{nd} order entities (either verbs, nouns or adjectives, i.e. the relation is type-persistent but can apply across PoSs). The only constraint is that the causing event should be dynamic, whereas the resulting situation can either be static or dynamic. Here are examples for adjectives:

```
illustrativo (= che serve ad illustrare) (illustrative)
    HAS_HYPERONYM
                           atto, adatto
                                                      (suitable
for)
         CAUSES non-factive, intention
         or
         FOR PURPOSE OF
                                   illustrare
                   (to illustrate)
difensivo (= atto a difendere)
                                   (defensive)
         HAS_HYPERONYM
                              atto, adatto
    (suitable for)
                       CAUSES non-factive, intention
         or
         FOR_PURPOSE_OF
                               difendere
    (to defend).
```

As said above, in IWN we are also going to encode some adverbs, derived from adjectives by means of the suffix *-mente*: e.g. *piacevolmente* (agreeably), derived from *piacevole* (agreeable), *dettagliatamente* (in detail), derived from *dettagliato* (detailed), etc. These will be directly linked to the corresponding adjectives by using the DERIVATION relation (cf. Table 2), while will turn out to have indirect links to all the synsets related with the corresponding adjectives (e.g. *dettagliatamente* will be indirectly linked to the noun *dettaglio* – detail).

Table 3 below gives an overview of the main new relations identified and being encoded in IWN (for a complete overview of the relations being encoded see Alonge et al., in prep.)¹⁴.

3.3 The IWN Top Ontology

In the EWN TO 2nd order entities have been organized into two main classification schemes:

- *Situation Type*: the event-structure or *Aktionsart* (or lexical aspect) of a situation;
- *Situation Components*: the most salient semantic components that characterize situations.

The Situation Types provide a classification of 2^{nd} order entities in terms of the event-structure (or *Aktionsart*) of the situation they refer to: a basic distinction was drawn between *Static* and *Dynamic*. The Situation Components can be viewed like the most salient semantic components of a concept. Situation Type represents disjoint features that cannot be combined, whereas it is possible to assign any combination of Situation Components to a word meaning. In table 4 the Top Concepts identified for 2^{nd} order entities are shown:

¹² A similar distinction is also made within the SIMPLE EC project (LE-8346), whose goal is adding semantic information to the set of harmonized lexicons built within the PAROLE project for twelve European languages. Of course, the sub-classification of antonymy can also be used for nouns and verbs.

¹³ Furthermore, these relations being encoded between an adjectival synset and a nominal or verbal one are also useful to distinguish adjective classes as described by Dixon (1991), and reported in Sanfilippo et al. (1999). Indeed, such classes are often indicated by the nouns linked to adjectives.

¹⁴ As it was done for all the relations identified in EWN, we have built substition tests or diagnostic frames based on normality judgements (cf. Cruse, 1986). Inserting two words in the test sentences built evokes a 'normality' 'abnormality' judgement on the basis of which each relation can be determined. These tests are used by encoders both to verify the existence of relations between synsets and to encode them in a consistent way (cf. Alonge et al., in prep.).

Relation	Semantic Orders Linked	Examples
COMPL_ANTONYM	1°/1°, 2°/2°, 3°/3	alive/dead
GRAD_ANTONYM	1°/1°, 2°/2°, 3°/3	cold/hot
RESULTS_IN/IS_RESULT_OF	2°/2°	to kill/to die
FOR_PURPOSE_OF/IS_PURPOSE_OF	2°/2°	to search/to find
IS_MEANS_FOR/HAS_MEANS	2°/2°	heat/to distill
LIABLE_TO/HAS_LIABILITY	2°/2°	triable/to judge
PERTAINS_TO/HAS_PERTAINED	2°/1° and viceversa	atomic/atom
IS_A_VALUE_OF/HAS_VALUE	2°/2°	tall/stature

Table 5. New relations being encoded in twive

2 ND Order Entity	
SITUATION TYPE	
Dynamic	
BoundedEvent	
UnboundedEvent	
Static	
Property	
Relation	
SITUATION COMPONENT	
Cause	
Communication	
Condition	
Existence	
Experience	
Location	
Manner	
Mental	
Modal	
Physical	
Possession	
Purpose	
Quantity	
Social	
Time	

Table 4: EWN 2nd order Top Concepts

In order to be able to draw generalizations also on adjective meanings by using the TO, we partially modified this scheme. First of all, we moved the PROPERTY and RELATION nodes under the SITUATION COMPONENT node. This was done for two interconnected reasons: first of all because this distinction is not directly linked to Aktionsart (lexical aspect), while the distinctions under SITUATION TYPE are Aktionsart distinctions, i.e. they are connected with the "the procedural characteristics (i.e. the 'phasal structure', 'time extension' and 'manner of development') ascribed to any given situation referred to by a verb phrase" (Bache, 1982:70)¹⁵. Secondly, adjectives may refer to PROPERTIES or RELATIONS, but they may be either stative or not (cfr. e.g. Lakoff, 1966; Quirk et al., 1985; Peters, Peters & Gaizauskas, 1999). Thus, in our system it has to be possible to specify e.g. that an adjective expresses a PROPERTY while being DYNAMIC. In any case, since many adjectives may have both a DYNAMIC sense

and a STATIC one, we have also the possibility to underspecify this information by linking adjectives directly to the SITUATION TYPE node.

Adjectives may indicate many different types of properties: temporal (passeggiata mattutina - morning walk), psychological (canzone triste - sad song), social (uomo ricco - rich man), physical (superficie legnosa wooden surface), physiological (bambino magro - thin child), perceptive (minestra calda - hot soup), quantitative (magra ricompensa - poor reward) and intensity properties (vino forte - strong wine). In the EWN TO there are already nodes which may be used to represent these distinctions (TIME, MENTAL, SOCIAL, PHYSICAL, QUANTITY) but we needed to better specify or also add some features. For example, we have added, under the already present node PHYSICAL, the node MATERIAL, to represent, among others, some Italian adjectives ending in -oso (for example *legnoso* - wooden, *acquoso* - watery) which indicate the property of containing a certain material. Moreover, we added the node PHYSIOLOGICAL (to classify adjectives corresponding to *tired*, *hungry*, *sick*, etc.) under PHYSICAL. For adjectives denoting intensity, we then added the node INTENSITY directly under the SITUATION COMPONENT node.

One of the main problem we had was that no Top Concept in the EWN TO could be used to classify reference-modifying adjectives (cf. Bolinger, 1967; Chierchia & McConnel-Ginet, 1990 name them *intensional adjectives*): i.e. adjectives like *former, future, present.* These are a very particular kind of adjectives, because they do not indicate a property of the referent of the noun they modify. So, aiming at showing the distinction between referent-modifiers and referencemodifiers, we created two new Top Concepts under the node PROPERTY: ATTRIBUTE and FUNCTIONAL, where the latter can be used for reference-modifying adjectives.

A particular case of functional adjectives are the 'argumental' ones. They introduce a comparison between different entities (e.g., *simile* - similar, *diverso* - different, etc.). A comparison presupposes a relation so these adjectives can be linked to both PROPERTY and RELATION. Since in the EWN TO these two Top Concepts were two different kinds of SITUATION TYPE, they were mutually exclusive; now, in the IWN revised TO they can be conjoined.

In table 5 the IWN Top Concepts for 2nd order entities are reported:

¹⁵ Of course, all 2nd order entities (and therefore also nouns or adjectives) can be classified according to their *Aktionsart*.

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

Table 5: IWN 2nd order Top Concepts

4. The Terminological Subset

IWN will include a terminological subset in the economic domain ('ECO-IWN'), which will be modeled according to the design principles of the generic wordnet. As a consequence ECO-IWN will make use of Italian synsets and of the internal relations among them; there will also be an economic ILI structured as the ILI of the generic wordnet, and a set of equivalence relations to connect the terminological synsets to the ILI economy synsets. In turn, the ILI economic synsets will be linked to proper concepts both of the Top Ontology and of the Domain Ontology.

As far as the methodology for the construction of ECO-IWN is concerned, a modular approach has been adopted, which allows to develop the specialized wordnet in parallel with the generic wordnet. Then, the integration between the two resources is made possible by means of a number of so-called 'plug-in' relations, which allows the two wordnets to be used jointly. A 'plug-in' relation connects a terminological sub-hierarchy (represented by its root node) to a node of the generic wordnet. When a plug-in relation is active, the effect is that, starting from the plug-in node, the terminological wordnet is preferred over the generic one. More in detail, a plug-in relation between a terminological node 'T-node' and a generic 'Gnode', implies that a new node (i.e. a PLUG-node) is created, which substitutes both T-node and G-node, with the following properties: i) all downward and horizontal

relations¹⁶ are taken from the terminological node; ii) all upward relations are taken from the generic node. This means that the new node inherits from the generic wordnet all the relations from which the node itself depends (for instance its hyperonyms); while it inherits from the specialized wordnet all the relations from which other nodes depend (for instance the hyponyms).

As for the connection of the terminological wordnet with the generic wordnet, one of the main problem is that inconsistencies may occur between the to resources, due to different points of view adopted (e.g. 'naive knowledge' versus 'expert knowledge'). For instance, in a generic wordnet (WN1.6) we find economic growth, which inherits from economic process, but a possible synonym such as *economic development* is not defined; instead, we find the economic sense of development (development #2), which inherits from process. A similar situation should be rearranged in the specialized wordnet, where the similarities between the two senses should be emphasized, first by adding economic development and then by connecting it to the generic development #2. To deal with similar cases a methodology making use of plug-in relations has been developed, which implements a shadowing mechanism of wordnet areas.

One of the main benefits of the plug-in approach is that information can be accessed in a modular way. In particular, once the plug-in relations have been defined, the user can decide among different views on the resource: plug-in relations can easily be made active, when a specialized view is preferred, or made inactive, if for the exploration of the resource a generic view is preferred. A second benefit is that a modular approach makes it possible to reuse already existent terminological resources, supposed that their model is wordnet compliant. In addition, given that both the generic and specialized resources share the same data structure, the IWN architecture allows each specialized wordnet to be used separately from the generic wordnet.

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.
- Alonge, A., Bertagna F., Calzolari N., Magnini B., Roventini A., in prep. ItalWordNet: a Large Lexical Resource for the Automatic Treatment of Italian.
- Bache, C., 1982. Aspect and Aktionsart: towards a Semantic Distinction. *Journal of Linguistics*, 18, 1.
- Bindi, R., Monachini M. & Orsolini P., 1989. Italian Reference Corpus, Istituto di Linguistica Computazionale del CNR, ILC-CNR-3, Pisa.
- Bolinger, D., 1967. Adjectives in English: Attribution and Predication. *Lingua*, 18, 1-34.
- Calzolari, N., Ceccotti, M. L., Roventini A., 1983. Documentazione sui tre nastri contenenti il DMI, Istituto di Linguistica Computazionale del CNR, ILC-DMI-2, Pisa.

¹⁶ The classification of wordnet relations based on their direction (upward, downward, horizontal) is described in Hirst & St-Onge (1998).

- Chierchia, G. & McConnel-Ginet S., 1990. An *Introduction to Semantics*. Cambridge, The MIT Press.
- Cruse, D. A., 1986. *Lexical Semantics*. Cambridge, Cambridge University Press.
- Dixon, R. M. W., 1991. A new Approach to English Grammar on Semantic Principles. Oxford, Clarendon Press.
- Gonzalo, J., Verdejo F., Peters C., Calzolari N., 1998. Applying EuroWordNet to Cross-Language Text Retrieval. In: Ide N., Greenstein D., Vossen P. (eds.), Special Issue on EuroWordNet. Computers and the Humanities, Vol. 32, Nos. 2-3, 185-207.
- Hirst, G. & St-Onge D., 1998. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In Fellbaum, C. (ed.) *WordNet, An Electronic Lexical Database*. Cambridge, The MIT Press.
- Lakoff, G., 1966. Stative Adjectives and Verbs in English. Computation Laboratory, Harvard University Report No. NSF-17.
- 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.

- Peters, I., Peters W. and Gaizauskas R., 1999. The Representation of Adjectives in SIMPLE. Ms.
- Quirk, R., Greenbaum S., Leech G. Svartvik J., 1985. A Comprehensive Grammar of the English Language. London, Longman.
- 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., Calzolari N., Peters C., Bertagna F., 1998. Building a Semantic Network for Italian Using Existing Lexical Resources. In: Antonio Rubio et al. (eds.), *First International Conference on Language Resources & Evaluation Proceedings*, Granada 28-30, 377-383.
- 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*, <u>http://www.hum.uva.nl/~ewn</u>.