The Lexico-semantic Annotation of an Italian Treebank

Nadia Mana*[*] and Ornella Corazzari[†]

[*]ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
via Sommarive 18, 38050 Povo - Trento, Italy
mana@itc.it

[†]CPR, Consorzio Pisa Ricerche
p.za A. D'Ancona 1, 56100 Pisa, Italy
corazzar@ilc.pi.cnr.it

Abstract

Corpora annotated at semantic level play a crucial role both in research and in applicative contexts in which systems of natural language processing are studied and developed. In this paper we present the lexico-semantic annotation of an Italian treebank, a first attempt to recover the lack of such resource for Italian. We will describe the annotation realized, focusing on the methodology followed, the results achieved, and possible further work and applications.

1. Introduction

Corpora annotated at semantic level play a key role both for the training and for the evaluation of applications in natural language processing (NLP). In particular, such corpora are crucial prerequisities for the automatic acquisition of linguistic knowledge, the retrieval and extraction of information from texts, and the testing and tuning of word sense disambiguation systems. In the same time corpora annotated are a useful testbed to evaluate the adequacy of the semantic lexicons and a good repository of corpus examples to attest the word senses. Finally, they turn out useful also for teaching purpose as source of examples, helpful to learn a language. In short, corpora annotated at semantic level and more in general at different annotation levels (also morpho-syntactic and syntactic) are extremely important both from the theoretical and the applicative point of view.

In order to recorver the lack of such resources for Italian, a first attempt has been realized within the national project SI-TAL[1] (Sistema Integrato per il Trattamento Automatico della Lingua - lit. 'Integrated System for the NLP'), oriented to build the first Italian Treebank with a multi-level annotation[2].

The lexico-semantic annotation described in this paper corresponds to the third annotation level of this treebank. This annotation has taken advantage of two previous experiments of semantic tagging performed at ILC (Istituto di Linguistica Computazionale in Pisa) in the framework of the SENSEVAL project (Calzolari and Corazzari, 2000) and of the development activity of an ELSNET resource (Corazzari and Monachini, 1995). In addition, the comparison with other treebanks, such as the Penn Treebank (Marcus et al., 1993), the Treebank for French (Abeillé et al., 2000), the Spanish Treebank (Moreno et al., 2000), has been of utmost importance.
In the next sections we will describe the lexico-semantic annotation realized within the SI-TAL project. In particular, we will focus on the methodology and process of annotation (Sec. 2.), and the results achieved within the project (Sec. 3.). Finally we will present some possible further works and applications (Sec. 4.).

2. Lexico-semantic Annotation

The lexico-semantic annotation described in this paper is basically a sense tagging (Kokkinakis and Kokkinakis, 1999) of lexical heads according to a reference lexical resource, called ItalWordNet (Roventini et al., 2000; Alonge et al., 2001). That resource includes a generic (i.e. domain independent) lexicon and a specialized (i.e. domain dependent) lexicon – specifically an economic one. The two lexical resources have been developed separately, but an integrated consultation is allowed by an integration procedure based on the definition of *plug-in* relations[3], as described in (Magnini and Speranza, 2001).

The lexico-semantic annotation consists basically in the assignment of semantic tags, expressed in terms of attribute/value pairs. For each lemma the annotator indicates the appropriate number of sense, reported in ItalWordNet, according to the specific context given by the corpus sentence containing that lemma. However, the annotation is more than a mere list of sense instantiations present in the reference lexical resource. First of all, since ItalWordNet has a taxonomical organization of word senses, the sense number assignment to the corpus words is an implicit assignment of the semantic types of ontology. Secondly, the

[1] The project, funded by MURST (the Italian Ministry of University and Scientific Research) and coordinated by CPR (Consorzio Pisa Ricerche), is a joint enterprise whose aim was the production of tools and resources for Italian NLP. Many research centres and institutions have been involved, in charge of different parts of the project : ILC-CNR/CPR and Synthema in Pisa, Venezia University/CVR in Venice, "Tor Vergata" University/CERTIA in Rome and ITC-irst in Trento.

[2] The SI-TAL treebank has a multi-level structure: (i) morphosyntactic level; (ii) syntactic level (with both costituency and functional annotation); (iii) lexico-semantic level. For more details, see (Montemagni et al., 2000).

[3] This *plug-in* approach is realized by a set of procedures that allows an integrated access of the two resources, such that sense overlappings are merged and conflict situations are properly managed

annotation has been enriched with additional information, in order to mark appropriately the following cases:

- **proper names** of:
  - persons (e.g. *Dante Alighieri*, *Agnelli*, *Carlo Maria Giulini*)
  - institutions or companies (e.g. *Arnoldo Mondadori*, *Dhl International*, *Omnitel spa*)
  - artifacts (e.g. *Fiat Brava*, *Windows 98*, *Nokia 3310*)
  - locations (e.g. *Milano*, *Gran Bretagna*, *Africa*)
- **alterations**, in case of units semantically modified through evaluative suffixations. For example, *libricino* (’small book’). In this case the reference lemma is `libro` and the value `dim` is assigned to the attribute `alter`; this means that the lemma is used with an alteration, namely a diminutive.
- **figurative uses**, such as metaphors (e.g. *essere una lumaca*, lit. ’to be a snail’ to mean ’to be slow like a snail’) or metonymies (e.g. *essere fedele alla bandiera*, lit. ’to be loyal to the flag’ to mean ’to be loyal to own country’ - country represented by the flag), lexicalized and non lexicalized.
- **idiomatic expressions**, for example *pagare a caro prezzo* (lit. ’to pay dearly for’) or *essere il fanalino di coda* (lit. ’to be the tail light’)
- **support verbs**, for example *fare effetto* (lit. ’to take effect’) or *entrare in vigore* (lit. ’to come into effect’), etc.
- **neologisms**, typical of specific domains. For example, expressions coming from English terms used in the computer field (e.g. *scannerizzazione* from the English ’scanner’ to indicate the operation of scannering) or expressions coming from the political or economic field (e.g. *cigiellino* to indicate a member of the CGIL union or *leghista* to indicate a member of the Lega Nord party.)
- semantically complex units indicating **titles**. In particular, three kinds of titles have been distinguished:
  - i) title of written texts, such as books, newspapers, magazines, etc., (e.g. *I Promessi Sposi*, *Gazzetta Ufficiale, Micro Mega*), marked as `semiotico`;
  - ii) title of programs, movies, etc. (e.g. *Il Festival di Sanremo*, *La vita è bella*), marked as `spettacolo`;
  - iii) generic titles of conferences, festivals, celebrations, events (e.g. *Il Salone del Libro*, *La settimana in Bulgaria*), marked as `tipotit`.

In the Table 1 the set of the main attribute/value pairs used during the annotation is reported.

Attribute	Value
lemma	the lemma annotated
pos	the part of speech in case of multiword expression
dbref	database name of the reference lexical resource (*gen* for ’generic’ and *eco* for ’economic’)
numero-senso	the sense number of the semantic unit according to the reference lexical resource
alterazione	kind of alteration in case of units semantically modified through evaluative suffixation
figurato	kind of figurative uses
nome-proprio	proper names of persons, institutions, artifacts, locations
tipousc	units semantically complex indicating idioms, compounds, support verb constructions, etc
tipolemma	the lemma type in case of terms in dialect or neologisms
tipotit	used specifically in correspondence of the polilexical expressions indicating titles, to distinguish titles of written texts , programs and generic titles
nota	to signal doubts of sense interpretation
commento	widely used by the annotators to signal cases of missing senses, missing lemma, foreign terms, doubts, part of speech errors, etc.

Table 1: Set of attribute/value pairs

2.1. Annotated Corpus

The corpus of this Italian treebank consists of two parts:

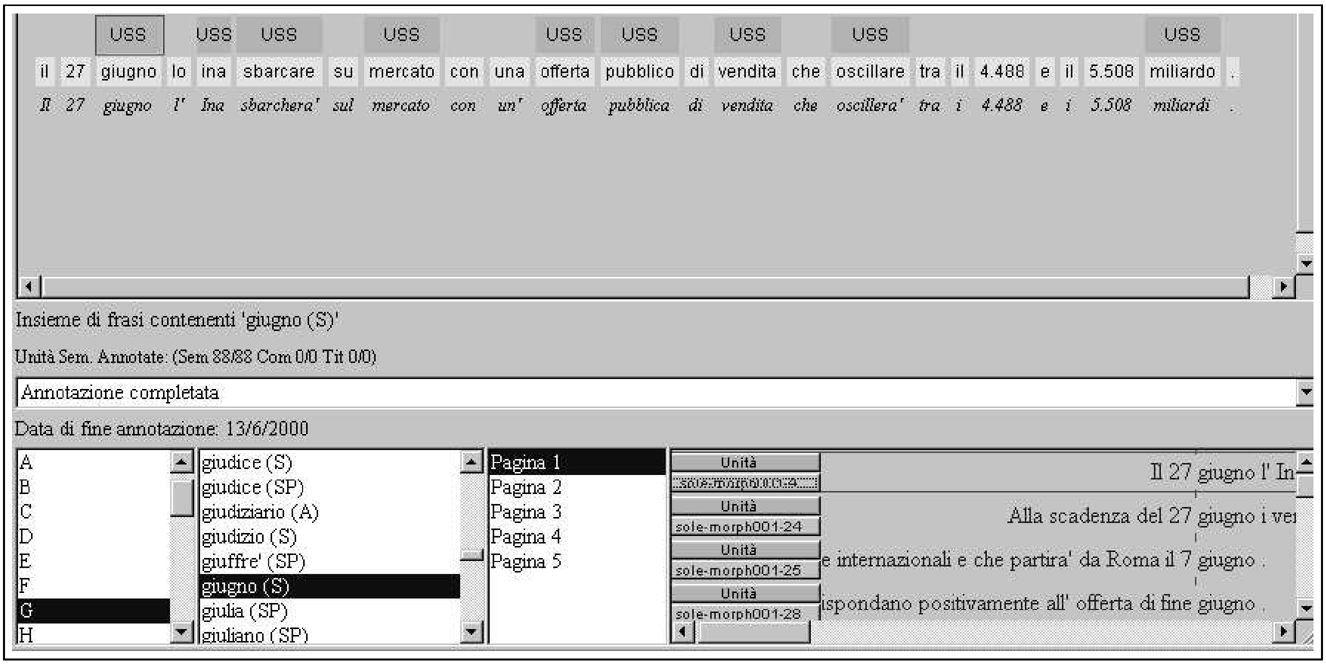
- a ”balanced corpus”[4], composed by articles from Italian newspapers and magazines, covering different topics - economy, politics, sport, culture, science, etc.
- a ”specialized corpus”, composed by economic and financial articles taken from *”Il Sole 24 Ore”*.

The size of the treebank corpus is about 300,000 word tokens (content and function words), but only a part has been annotated at the lexico-semantic level. In particular, about 80,000 tokens (intended here as content words: nouns, verbs and adjectives), of which 56,000 of the ”balanced corpus” and 24,000 of the ”specialized corpus”.

2.2. Annotation Methodology

Before starting the annotation, the Treebank staff defined the technical guidelines for the lexico-semantic annotation, through many meetings and discusssions, taking into

[4] As reported in (Teubert, 1995; Macleod et al., 1998), a ”balanced corpus” is a type of corpus composed according to parameters such as text type, genre or domain.



il 27 giugno lo ina sbarcare su mercato con una offerta pubblico di vendita che oscillare tra il 4.488 e il 5.508 miliardo .

Il 27 giugno l' Ina sbarchera' sul mercato con un' offerta pubblica di vendita che oscillera' tra i 4.488 e i 5.508 miliardi .

Insieme di frasi contenenti 'giugno (S)'

Unità Sem. Annotate: (Sem 88/88 Com 0/0 Tit 0/0)

Annotazione completata

Data di fine annotazione: 13/6/2000

A
B
C
D
E
F
G
H

giudice (S)
giudice (SP)
giudiziario (A)
giudizio (S)
giuffre' (SP)
giugno (S)
giulia (SP)
giuliano (SP)

Pagina 1
Pagina 2
Pagina 3
Pagina 4
Pagina 5

[illegible] Il 27 giugno l' In
sole-morph001-24 Alla scadenza del 27 giugno i ver
sole-morph001-25 e internazionali e che partira' da Roma il 7 giugno .
sole-morph001-28 ispondano positivamente all' offerta di fine giugno .

Figure 1: A view of the main GesTALt window

account other relevant works in the literature on semantic annotation for other languages.

Three kinds of semantic units have been annotated:

1. **uss** (unità semantica semplice - ’simple semantic unit’): sense units that correspond to single lexical items (e.g. *prezzo* ’price’; *politica* ’politics’; *palazzo* ’palace’);

2. **usc** (unità semantica complessa - ’complex semantic unit’): semantically complex units expressed in terms of multi-words, as compounds (e.g. *consiglio di amministrazione* lit. ’board of directors’), support verb constructions (e.g. *fare paragoni* lit. ’to make comparisons’ in the meaning of ’to compare’), idioms (e.g. ’prendere due piccioni con una fava’ lit. ’to take two pigeons with a broad bean’ corresponding to the English expression ’to kill two birds with one stone’);

3. **ust** (unità semantica titolo - ’title semantic unit’): sense units that correspond to titles of newspapers (e.g. *La Repubblica*, *Il Sole 24 Ore*, *Il Corriere della Sera*, etc.), books (e.g. *Il nome della rosa*, *La Divina Commedia*, *Pinocchio*, etc.), operas (e.g. *Le nozze di Figaro*, *La Cenerentola*, *Don Giovanni*).

A different treatment has been defined for the polilexical expressions indicating titles: whereas **uss** and **usc** have been annotated at one level only, all **ust** (e.g. the opera *Le nozze di Figaro*) have received a two-level annotation (one at the level of individual components and one at the level of the whole unit). With reference to the example above, according to this strategy, *nozze* and *Figaro* have been annotated individually as **uss** and the whole string *Le nozze di Figaro* as **ust**. In this way, we do not preclude the possibility of processing both on the titles and their internal components[5].

2.3. Annotation Procedure

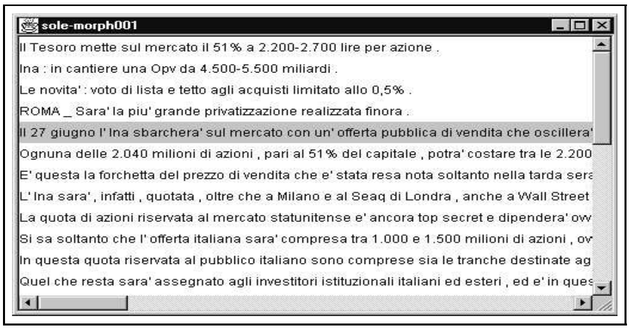
Whereas the syntactic annotations have been performed by a word-by-word approach, the lexico-semantic annotation process has been carried out by lemmas: chosen a lemma and verified its possible senses in ItalWordnet, all its occurrences in the corpus have been annotated. In this way more homogeneous and coherent annotations have been guaranteed. This annotation was mostly accomplished by following the criterion of relative frequency within each corpus, i.e. annotating most frequent lemmas first, and so on. The annotation process started from the lemmas shared between the balanced and financial corpora. These lemmas, decided in accordance with the requirements of the ItalWordNet lexicographers, were chosen in order to allow for an higher stability of the reference lexical resource – the meaning of such lemmas were thoroughly verified before the annotation took place.

Instead of using a semi-automatic approach as described, for example, by Erdmann et al. (2000), the semantic annotation has been carried out manually, with the support of a tool (GesTALt[6]) specifically developed for building the treebank. As shown in the Figure 1, the access to the lemma listing, ordered alphabetically, is simplify by an appropriate graphical interface.

When the annotator selects an alphabet letter (e.g. the ’g’ letter), the tool shows all lemma starting with that letter. When selecting a lemma (e.g. *giugno*), all occuren-

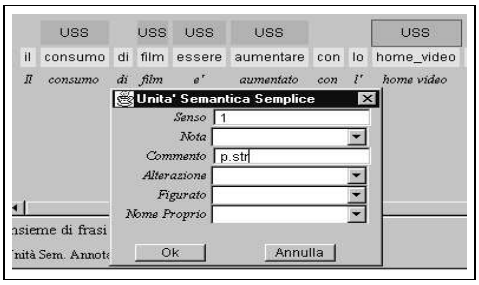
[5] That could be useful, for example, for Information Retrieval tasks, where the queries can regard whole titles but also subparts.

[6] The GesTALt tool has been designed and developed at CERTIA - Centro per la Ricerca, Sviluppo, Formazione nelle Tecnologie e Applicazioni- in Rome, in collaboration with Tor Vergata University.



Il Tesoro mette sul mercato il 51% a 2.200-2.700 lire per azione .
Ina : in cantiere una Opv da 4.500-5.500 miliardi .
Le novita' : voto di lista e tetto agli acquisti limitato allo 0,5% .
ROMA _ Sara' la piu' grande privatizzazione realizzata finora .
Il 27 giugno l' Ina sbarchera' sul mercato con un' offerta pubblica di vendita che oscillera'
Ognuna delle 2.040 milioni di azioni , pari al 51% del capitale , potra' costare tra le 2.200
E' questa la forchetta del prezzo di vendita che e' stata resa nota soltanto nella tarda sera
L' Ina sara' , infatti , quotata , oltre che a Milano e al Seaq di Londra , anche a Wall Street
La quota di azioni riservata al mercato statunitense e' ancora top secret e dipendera' ovv
Si sa soltanto che l' offerta italiana sara' compresa tra 1.000 e 1.500 milioni di azioni , ov
In questa quota riservata al pubblico italiano sono comprese sia le tranche destinate ag
Quel che resta sara' assegnato agli investitori istituzionali italiani ed esteri , ed e' in ques

Figure 2: Article from *’Sole 24 Ore’*



Senso: 1
Nota:
Commento: p.str
Alterazione:
Figurato:
Nome Proprio:

Figure 3: Annotation example of a USS

cies of that lemma in the corpus are shown, splitted in one o more pages (20 occurrencies, and so 20 sentences, per page). In the right part of the tool window, the news number and the sentence containing that lemma are reported. Clicking on the news number (e.g. *sole-morph001-4*), the whole sentence is visualized on the top of the window. In this way the annotator can see the complete sentence context in which that word is positioned. If the sentence context is not enough to understand in which sense the lemma is used, the annotator can look the whole news over simply clicking on the sentence on the right part of the window, at the bottom. In the news text the sentence containing the lemma in examination is easly detectable because it is highlighted in yellow (see Figure 2).

At this stage, the real annotation phase starts according to the following steps (not rigorously in this order):

- to search for the specific lemma in the *ItalWordNet* linguistic resource, and to evaluate the corresponding allowable meanings.

- to evaluate the context of each word occurrence of the given lemma, in order to obtain a context-dependent sense for the lemma itself. The context is defined by the sentence including the given lemma. In practice, this step consists in reading and analyzing the single lemma with respect to the phrase containing it. Whenever necessary, the previous step is repeated more times increasing the context in order to capture the lemma sense – the reference text may include the corresponding sentence or the paragraph or even the whole article.

- to select the text portion to be annotated. It can be the bare lemma, but in other cases it may turn out to be a compound semantic unit. Consider that certain polilexical expressions were already provided in input (e.g. *box office*, *ad hoc*, *in funzione di*), and managed as compounds from the very beginning, throughout every annotation level. Other cases, featuring a stronger semantic-lexical component (e.g. support verb constructions, idioms, names of persons, companies, manufactures), are explicitly created only at the present level.

- to assign a proper meaning (namely sense number) to the semantic unit under consideration (either simple, complex, or title), among those considered by *ItalWordNet*. If a given lemma or a specific sense is missing in the reference lexical resource, a discussion takes place among the annotation staff members and, if required, the problem is issued to the *ItalWordNet* staff. If a specific lemma does not appear in the lexicons (neither generic, nor specialistic) due to its limited suitability, the conventional value `no` is assigned to the sense attribute. On the contrary, if the lemma has more senses attested in the reference lexical resource and allowable in the specific context[7], the annotator reports all the good sense numbers, concatenating them with the symbol & (e.g. `1&3`).

- to enrich the annotation with additional information in order to mark proper nouns, idioms, metaphors, support verbs, etc., as agreed within the treebank staff.

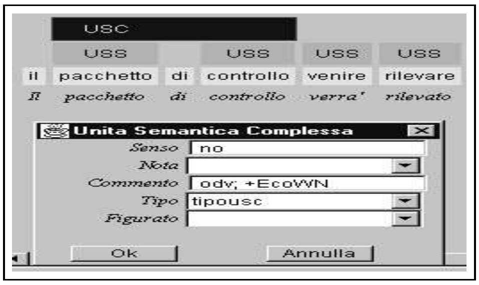
2.4. Some Examples of Annotation

The annotation methodology is here illustrated by some examples, one for each kind of semantic units.

First of all, the Figure 3 shows an example of **uss** annotation. In the example the target word is *home video* (polilexical expression, already provided as a single unit at the morpho-syntactic level) in the context given by the sentence *’il consumo di film è aumentato con l’home video’* (lit. ’the film consumption increased by means of the home video’). As the annotation window shows, the value `1` has been assigned to the attribute `senso` because the lemma is included into *ItalWordNet* resource, as foreign term in usage in the Italian language. The annotator has pointed out that the target word is a foreign word specifying `p.str` (abbreviation of *parola straniera* ’foreign word’) in the slot `Commento`.

Secondly, in Fig. 4 an example of **usc** annotation is shown. The annotated word is ’*pacchetto di controllo*’ (lit. ’shareholding’), a multi-word expression, typical of the economical domain. As shown in the annotation window, the annotator has built the **usc** (see the dark box marked `USC`, above the `USS` boxes which are empty). Then the value `no` has been assigned to the attribute `senso` because *ItalWordNet* did not include that expression (neither

[7] Consider for example the sentence “*un medico mi ha detto che c’era un problema*” lit. ’a doctor said me that there was a problem’. As attested in ItalWordNet, the verb *dire* can be used with 7 different senses. Among these, for the sentence above both the senses 1 (*dire, enunciare, proferire* - ’to say’) and 3 (dire, far sapere - ’to inform’) are good.



USC

USS USS USS USS

il pacchetto di controllo venire rilevare

Il pacchetto di controllo verra' rilevato

Unita Semantica Complessa

Senso no

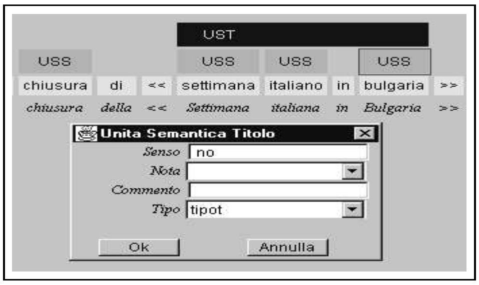
Nota

Commento odv; +EcoWN

Tipo tipousc

Figurato

Figure 4: Annotation example of a USC



UST

USS USS USS USS

chiusura di << settimana italiano in bulgaria >>

chiusura della << Settimana italiana in Bulgaria >>

Unita Semantica Titolo

Senso no

Nota

Commento

Tipo tipot

Figure 5: Annotation example of a UST

in the generic resource nor in the specialized one). However, since the expression is common in the economic language, the annotator has suggested to add it into the lexical reference resource. In particolar, he/she uses the field `Commento` to mark that the lemma will be object of evaluation (`odv`) on the ItalWordNet side and to suggest an inclusion in the economic lexical resource (`+EcoWN`). In the same time, the field `Tipo` is used to specify that the lemma is a compound unit (`tipousc`).

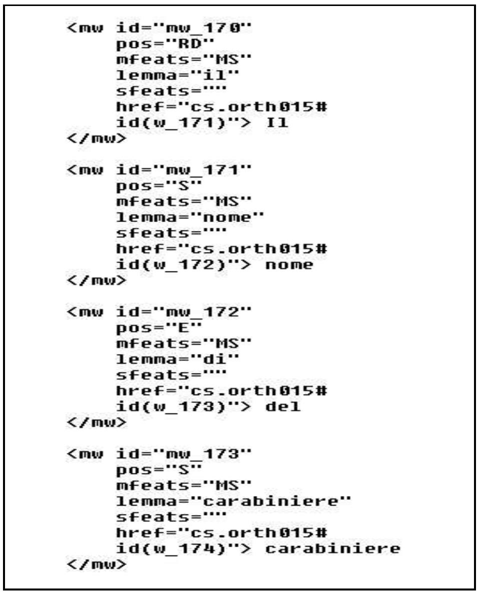
Finally, an example of **ust** annotation (Figure 5). The annotated title is '*Settimana in Bulgaria*'. It has been annotated at two levels: at the first one, the single constituting units (*settimana* and *Bulgaria*) have been tagged according to their individual sense, as reported in *ItalWordNet*. At the second level, the title has been treated as a single unit: the annotator has built the complex unit (as the dark box named `UST` shows) and marked with `tipo=`*`tipotit`* to indicate that it is a generic title, referring to an event. Obviously, at **ust** level the sense number information are missing and so the attribute `senso` is tagged as `no`.

3. Annotation Results

The final output of the lexico-semantic annotation level has been coded in XML format. In order to avoid redundancy, the lexico-semantic annotation level is directly linked to the corresponding annotation at the morpho-syntactic one. For example, given the sentence (extracted from *Il Corriere della Sera*):

> *Il nome del carabiniere (i corazzieri sono infatti un reparto speciale dell'Arma) è "top segret", ma la vicenda è stata denunciata alla magistratura e il procuratore militare Antonino Intelisano ha aperto un'inchiesta per il reato di offesa all'onore del Presidente della Repubblica.*

there is a XML representation of both the morpho-syntactic annotation (see Figure 6) and the corresponding lexico-semantic annotation (Figure 7).



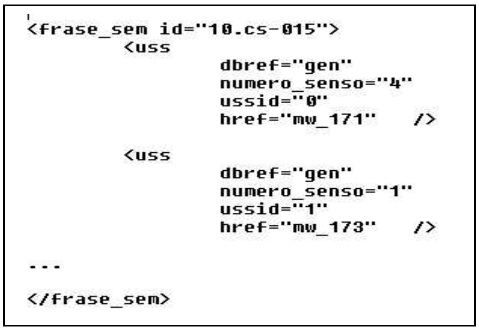
```
<mw id="mw_170"
    pos="RD"
    mfeats="MS"
    lemma="il"
    sfeats=""
    href="cs.orth015#
    id(w_171)"> Il
</mw>

<mw id="mw_171"
    pos="S"
    mfeats="MS"
    lemma="nome"
    sfeats=""
    href="cs.orth015#
    id(w_172)"> nome
</mw>

<mw id="mw_172"
    pos="E"
    mfeats="MS"
    lemma="di"
    sfeats=""
    href="cs.orth015#
    id(w_173)"> del
</mw>

<mw id="mw_173"
    pos="S"
    mfeats="MS"
    lemma="carabiniere"
    sfeats=""
    href="cs.orth015#
    id(w_174)"> carabiniere
</mw>
```

Figure 6: XML representation of a text example annotated morpho-syntactically



```
<frase_sem id="10.cs-015">
        <uss
                dbref="gen"
                numero_senso="4"
                ussid="0"
                href="mw_171"   />

        <uss
                dbref="gen"
                numero_senso="1"
                ussid="1"
                href="mw_173"   />

...

</frase_sem>
```

Figure 7: XML representation of a text example annotated at lexico-semantic level

At the semantic level the annotated lemmas (e.g. *nome* and *carabiniere*) are not esplicitly included in the XML representation, but they are simply linked to the morpho-syntactic level by the reference identifier (i.e. `mw_171` and `mw_173` respectively).

The Table 2 summarizes the results concerning the lexico-semantic annotation, achieved within the SI-TAL project, in terms of number of annotated semantic units. Globally 81,236 content word occurencies have been annotated, so distributed: 65,141 **uss** - 4,548 **usc** - 283 **ust**.

4. Possible Further Work and Applications

From many points of view, the lexico-semantic annotation realized within the SI-TAL project is not a conclusive work. As for the other annotation levels, further work could be put in practice. First of all, refinements could be made, namely correction of errors and re-annotation of problematic cases (appropriately marked by the annotator during the annotation). Secondly, extensions could be provided,

Corpus	Source	USS	USC	UST	Total Coverage
Balanced	*La Repubblica*	30,993	2,240	174	39,730
Balanced	*Il Corriere della Sera*	12,621	981	76	16,368
Specialised	*Il-Sole-24-Ore*	21,527	1,327	26	25,138
Spec+Balanced	*Total*	65,141	4,548	283	**81,236**

Table 2: Annotated semantic units

adding new texts in order to increase the size, to improve the language coverage or to define new domains (e.g. medical domain). Then, also an evaluation of the annotation consistency could be realized. In such a case, a sample from each corpus (balanced and specialized) could be extracted and the relative annotations could be checked by one or more people. Also an evaluation of the inter-annotator agreement could be carried out. However, this operation would need more annotators of the same corpus[8]. Finally, a further work could regard a language analysis in terms of word frequency, sense frequency, compound word frequency, etc., with possible comparison between the generic and specific domain.

As far as applications of the treebank are concerned, it has already been used. Firstly, to tune an Italian-English machine translation system (PeTra[9]). Secondly to test sense disambiguation systems in the SENSEVAL-2 competition[10]. The lexico-semantic annotation of the treebank could be exploited also for training and tuning of NLP systems, such as Information Extraction, Question/Answering, Information Retrieval, Summarization.

5. Conclusion

In this paper we have described the lexico-semantic annotation of the Italian treebank within the SI-TAL project. The novelty of this treebank lies not in the annotation methodology but first of all in the fact that it is the first treebank for Italian of a considerable size. Secondly, another relevant feature is that the lexico-semantic annotation is not a mere list of sense instantations according to ItalWordNet, but it provides many useful additional information. Finally, also the multi-level structure (morpho-syntactic, syntactic and lexico-semantic levels) of the treebank represents a novel aspect. It is not the only treebank including syntactic and semantic level at the same time. For example, also the Prague Dependency Treebank (Bemova et al., 1999) includes both levels. However, the Italian treebank is distinguished by conceiving semantic annotation as a sense tagging of lexical heads. In this way, the requirements for corpus-based investigations are guaranteed: by linking the syntactic and semantic annotation layers, it is possible, for istance, to identify specific subcategorisation properties associated with a specific word sense or to get the semantic types associated with functional positions for a given predicate.

[8] Consider that, in order to guarantee a higher grade of consistency, all the lexico-semantic annotation has been carried out by two persons: one for the balanced corpus and the other one for the specialized corpus.

[9] PeTra is an application based on the 'Slot Grammar' formalism, defined by Michael McCord at IBM T.J. Watson Research Center (McCord, 1990). For more details on PeTra see (Fanciulli and Raffaelli, 2001)

[10] The purpose of SENSEVAL-2 is to evaluate the strengths and weaknesses of sense disambiguation systems with respect to different words, different varieties of language, and different languages. The competition has involved many systems for several languages. Among these, also Italian.

6. References

A. Abeillé, L. CLement, and A. Kinyon. 2000. Building a Treebank for French. In *Proceedings of LREC-2000*, pages 87–94, Athens, Greece.

A. Alonge, F. Bertagna, N. Calzolari, O. Corazzari, and A. Roventini. 2001. ItalWordNet: extending an existing resource for NLP. In *Proceedings of the NAACL-2001 Workshop "Wordnet and other Lexical Resources: Applications, Extensions and Customizations"*, Pittsburgh, PA.

A. Bemova, J. Hajic, B. Hladka, and J. Panemova. 1999. Syntactic Tagging of the Prague dependency Treebank. In *Proceedings of Treebank Workshop, Journee(s) ATALA sur les corpus annotes pour la syntaxe*, pages 87–94, Universite Paris, place Jussieeu 7, Paris, France.

N. Calzolari and O. Corazzari. 2000. Sensereval/Romanseval: the framework for Italian. *Computers and the Humanities*, 34(1-2):61–78.

O. Corazzari and M. Monachini. 1995. ELSNET: Italian Corpus Sample. ILC, Pisa, Italy.

M. Erdmann, A. Maedche, H. P. Schnurr, and S. Staab. 2000. From Manual to Semin-automatic Semantic Annotation: About Ontology-based Text Annotation Tool. In *Proceedings of the COLING Workshop on Semantic Annotation LINC-2000*, Luxembourg, August.

F. Fanciulli and R. Raffaelli. 2001. Utilizzo della risorsa Treebank per il tuning di un sistema di traduzione. In *Proceedings of the Workshop on 'LA Treebank Sintattico-Semantica dell'Italiano di SI-TAL*, Bari, Italy.

D. Kokkinakis and S. J. Kokkinakis. 1999. Sense-Tagging at the Cycle-Level Using GLDB. Goteborg University.

C. Macleod, R. Grishman, and A. Meyers. 1998. Dictionaries and Balanced Corpora: The interdependence of resources. Technical report, Department of Computer Science - New York University.

B. Magnini and M. Speranza. 2001. Integrating Generic and Specilized Wordntes. In *RANLP: Recent Advances in Natural Language Processing*, pages 149–153, Tzigov Chark, Bulgaria.

M. Marcus, M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M.C. McCord. 1990. Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars. *Computer Science*, pages 118–145.

S. Montemagni, F. Barsotti, N. Calzolari, O. Corazzari, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, and R. Delmonte. 2000. Building the Italian Syntactic-Semantic Treebank. In *Building and using syntactically annotated corpora*. Kluwer, Dordrecht.

A. Moreno, A. Lopez Ruesga, S. Sanchez Leon, and F. Sanchez. 2000. Spanish Treebank: Specifications. In *Building and using syntactically annotated corpora*. Kluwer, Dordrecht.

A. Roventini, A. Alonge, F. Bertagna, B. Magnini, and N. Calzolari. 2000. ItalWordNet: a Large Semantic Database for Italian. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, pages 783–790, Athens, Greece.

W. Teubert. 1995. *Language Resources: The Foundations of a Pan-European Information Society*. Heike Rettig Ed.